CSIRO

Mathematical and Information Sciences

An Update on Estimating a CPUE Series for Southern Bluefin Tuna Using Enhanced Tree-based Modelling Methods

William N Venables[1], Peter J Toscas[1,3], Mark Bravington[1]
and Tom Polacheck[2]

August, 2003

Report Number:

2003/141

[1] CSIRO Mathematical and Information Sciences,
PO Box 120, Cleveland, Qld, 4163, AUSTRALIA

[2] CSIRO Marine Research, Hobart, AUSTRALIA

[3] Author for all correspondence. Peter.Toscas@csiro.au

CCSBT-ESC/0309/28

Table of Contents

List of Tables

List of Figures

Abstract

We present an update to a previous contribution that considered a tree-based method for estimating an index of abundance for the Southern Bluefin Tuna fishery using a non-parametric approach. Specifically, we consider tree-based models for *CPUE* data, enhanced by the technique known as bootstrap aggregation, or 'bagging'. We find it useful to estimate *expected effort* as a preliminary step. *CPUE* is then estimated conditional on effort, evaluated at the expected effort. The methods model zero and non-zero responses separately, for several reasons, but incidentally avoiding the need to adjust zero values artificially before taking logarithms. With strongly empirical models there is a danger that spurious relationships between effort and CPUE will be detected. This update goes some way to addressing that issue by a simulation study to show how the discovered relationships break down as the degree of spatial integrity in the data is artificially degraded.

Introduction

The Southern Bluefin Tuna (SBT) is a very mobile fish but not much is known about its movement. The little that is known suggests that the SBT has a very aggregated distribution and the location of concentrations of fish (based on catch rates) change with time, both within and between years (see Polacheck and Preece 2001). The spatial pattern of fishing effort also changes with time (see also Polacheck and Preece 2001), which may (or may not) reflect the changing spatial distribution of fish abundance.

Interpretation of *CPUE* data is complicated by the fact that fishers will naturally attempt to target their effort in regions with high density. Moreover, the spatial distribution of fish changes over time and the changes are likely to be at least partially a function of the overall abundance (e.g. density dependent habitat effects), which is the quantity being estimated. Thus, the distribution of effort in those squares fished does not constitute a "representative" sample and the estimation of the average density (including areas not fished) needs to take into account a bias and changing biases in the distribution of effort. Hearn and Polacheck (1998) present several indices of abundance for SBT based on bounds for the ability of fishers to target successfully those areas where SBT occur. At the one extreme it is assumed that fishers have no ability to target areas of higher density and that the areas with no effort have similar densities to those that have been fished. At the other extreme, it is assumed that fishers have perfect ability to target areas with SBT and thus areas with no effort have zero abundance. These scenarios were developed as a method for placing credible bounds on the relative density in the temporal/spatial strata in which there was no fishing effort. The two scenarios were seen as embedding two ends of a continuum in the relationship between the distribution of fishing effort and the underlying abundance of SBT (that is, the ability to target effort in high density areas.)

Other approaches have been adopted to estimate the *CPUE* in unfished squares. Thomas and Toscas (1997) and Toscas *et al*. (1998) considered a geostatistical approach (see Cressie 1993, Chapter 3, and SAS 1999, `PROC MIXED`) to the estimation of abundance in unfished squares. The abundance in these squares was estimated from the *CPUE* of adjacent squares in the same year and month. Toscas *et al*. (2001) extended the geostatistical approach to incorporate a temporal component. In this paper the abundance in unfished squares was estimated from the *CPUE* of adjacent squares in both time and space. Hearn and Polacheck (1998) use the general linear model to estimate the *CPUE* in unfished squares.

These parametric analyses suggest that there are significant higher order interactions between year, area and seasons for *CPUE*. Strongly factorial models with high order interactions are often a problem simply because of the large number of degrees of freedom they require, as often only a relatively small number of them are useful in the sense of adding to the predictive capacity of the model. The majority of degrees of freedom for such inflated models often transfer error from the data set into errors of prediction. For parametric models to work well we usually need very effective explanatory (or predictor) variables to be available. Moreover these explanatory variables have to be available at matching spatial and temporal scales to that of the observations themselves. At present we do not have explanatory variables that meet these requirements for other than a very short part of the history of the fishery.

The approach presented in this paper is an attempt to bypass these and other difficulties with enhanced regression and classification tree modelling techniques that appear to be able to generate surrogates for local predictors using simply spatial and temporal terms. This approach seems better able to capture abrupt changes in the response, and complex interactions between time and space covariates are in effect automatically generated as needed. The enhancement is bootstrap aggregation or 'bagging' (see Breiman 1996 and 1998). Watters and Deriso (2000) used regression trees to estimate indices of abundance from *CPUE* data for Bigeye tuna and Shono *et al.* (2001) used the same technique to estimate annual *CPUE* indices for SBT.

The method in this paper also avoids the need to adjust zero quantities in order to take logarithms, which we regard as a defect of most of the present approaches, (including some former ones of ours). We prefer to model the zero component of a distribution as a separate mixture component of the whole distribution and then to use standard techniques on the positive component, for which taking logarithms directly is feasible, if necessary. Since regression trees are fitted by a least squares process, it is necessary to transform the response in order to ensure reasonable homogeneity of error variance, and the log transformation seems appropriate for this in our cases.

The next section outlines the data used in the analyses and the statistical methodology. The section after that details the results and the paper concludes with a discussion.

Materials and methods

Data and definition of the fishery

Data were provided by the National Research Institute of Far Seas Fisheries (NRIFSF) within the context of the Commission for the Conservation of the Southern Bluefin Tuna (CCSBT) assessments. The term 'square' is used for a fixed 5° × 5° geographical region between 30° and 50° south and 340° to 185° east for which catch and effort data are available. The term 'cell' will normally be used for a square in a given month of the year but sometimes it will mean a specific square of a given month of a given year, as will be clear from the context. For this paper the fishery has been defined as the collection of cells that have received fishing effort in any of the years 1969 to 2000. Thus the extent of the fishery varies from month to month and for any given month consists of all squares that have been fished at least once in that month throughout the data record available. For each year data are available for the months April to September.

The weighting scheme used to calculate annual indices is the same as the scheme described in Hearn and Polacheck (1998). Each cell is given a weight that does not vary from year to year, but may vary between second and third quarters. Annual indices must take due account of the spatial distribution of effort and abundance. This is achieved by annual estimation of catch per unit effort averaged over the entire fishery, with weights that vary spatially in such a way as to reflect the spatial allocation of effort.

Modelling strategy

Our joint distribution of effort and catch may be described as follows:

1. The marginal distribution of effort, or hooks, (*H*), is a mixture of a discrete probability at zero and a continuous distribution with positive support. The probability that a cell has zero hooks allocated is modelled as a function of spatial and temporal covariates. The conditional mean log-number of hooks allocated, given that the allocation is positive, is also modelled as a function of the same covariates.
2. The conditional distribution of catch given effort ($C \mid H$) for a cell, is clearly zero with probability 1 if $H = 0$, but otherwise is also a two-component mixture, again with the discrete probability mass at zero and the conditional mean log-catch (given both $H > 0$ and $C > 0$), both functions of the expected log-hooks as well as of spatial and temporal covariates.
3. In both cases the variance of the positive component on the log scale is assumed to be constant.

A primary analysis on the log scale is needed because (a) most of the influences on the hooks and catch distributions are likely to be multiplicative and hence (b) the coefficient of variation is likely to be close to constant. The analysis method uses least squares, which if it is to be efficient must have a relatively constant variance in the response.

For fished cells, that is, for cells with non-zero effort assigned, the obvious estimator of *CPUE* is the ratio of catch to hooks. This is called the nominal *CPUE* for the cell. The smoothed estimate of *CPUE* is the ratio of the estimated mean catch given the applied effort to that applied effort. For cells with no effort assigned it seems reasonable to estimate potential *CPUE* by first estimating the mean number of hooks for the cell (which for any cell in the fishery ought to be positive, though in some cases will be very small) and then estimating the mean catch conditional on effort being applied at its expected value. This leaves open the case where the estimated mean number of hooks for a given cell is zero. This can happen with simple tree-model estimators due solely to the roughness of the model. Bagging, however, when applied sufficiently vigorously seems to provide non-zero mean estimates for hooks in every cell of the fishery, though indeed for some cells the estimate is very small. The definition of the fishery we have adopted here is consistent with that of previous studies. An alternative definition for the future may be that a cell belongs to the fishery only if its expected mean allocation of hooks is positive.

Finally an estimate of the *CPUE* abundance index is obtained by weighted averaging of local *CPUE* estimates.

Estimating the probability that for a given year a cell is fished is achieved by bagging classification trees (see Breiman, 1996). This involves forming *M* replicate data sets, each consisting of *N* randomly selected observations (with replacement) from the original data set – so called 'bootstrap samples' – and constructing a classification tree on each of the *M* replicate data sets. Large values of *M* take much more computational time, of course. For the analyses in this paper *M* was chosen to be 100. Note also that the bagging procedure also has a similar function to cross-validation and pruning for a single tree, so the control on the complexity of individual trees can afford to be fairly simple.

The bagged estimate of the probability of a cell being fished in a given year can then be calculated as the average of the estimated probabilities from the *M* bootstrapped replicate data sets. The tree model works with a deviance measure that is zero if and only if the final node is homogeneous, that is, it is a classification as opposed to a regression tree model. The explanatory variables used to estimate the probability that a cell is fished are year, month, latitude and longitude, with the first three considered to be nominal factors and the last numeric.

Estimating the expected number of hooks allocated to any cell in a year is a two-stage process. The first step is to estimate the conditional expectation of log(H) | $H > 0$, where H denotes 'the allocated hooks to a cell', by regression trees. (Note that upper case italics are used to represent random variables.) Bagging a regression tree is similar to bagging a classification tree, except that the bagging prediction is the average of the estimated conditional expectations from each of the *M* regression trees. For cell *i* this can be expressed as:

$$\mathrm{E}\left[\log\left(H_i\right) \mid H_i > 0\right] \cong \tfrac{1}{M}\sum_{j=1}^{M} \mathrm{BSE}_j\left[\log(H_i) \mid H_i > 0\right], \qquad (1)$$

where BSE_j denotes 'the estimate from the *j*th bootstrap sample of [...]' (and '$\cong$' is to be read 'is estimated by'). The second stage estimates the unconditional expected number of hooks deployed in a cell, which uses the result:

$$\begin{aligned}\mathrm{E}\left[H_i\right] &= \mathrm{P}\left(H_i > 0\right)\exp\left\{\mathrm{E}\left[\log\left(H_i\right) \mid H_i > 0\right] + \tfrac{1}{2}\sigma_{lh}^2\right\} + \mathrm{P}\left(H_i = 0\right)\times 0 \\ &= \mathrm{P}\left(H_i > 0\right)\exp\left\{\mathrm{E}\left[\log\left(H_i\right) \mid H_i > 0\right] + \tfrac{1}{2}\sigma_{lh}^2\right\},\end{aligned} \qquad (2)$$

where $\mathrm{P}\left(H_i > 0\right)$ uses the bagged estimate for the probability that cell *i* has effort and the quantity σ_{lh}^2 used in a simple bias correction for the back transformation of log hooks is estimated as:

$$\sigma_{lh}^2 \cong \frac{1}{n_h}\sum_{k=1}^{n_h}\left\{\log\left(h_k\right) - E\left[\log\left(H_k\right) \mid H_k > 0\right]\right\}^2,$$

where n_h is the number of observed cells with effort. (Here and elsewhere a lower-case version of a term is used to represent the observed value of the random variable.) The predictor variables used in the (bagged) tree models for the expectation E[log(H) | $H > 0$] include the terms used for the probability that a square was fished and the probability of effort itself.

Estimating the probability that there is a non-zero catch in a cell is another two-stage process. The first stage uses bagged classification trees to estimate the conditional probability that a cell has some non-zero catch of SBT given $H > 0$. The explanatory variables used to train the model are year, month, latitude, longitude, probability of effort, observed hooks, and both unconditional and conditional expected hooks. The explanatory variables year, month and latitude are once again treated as factors and longitude as a numeric variable.

Estimating the expected number caught in a cell is also a two-stage process. Firstly, regression trees are used to estimate the conditional expectation $\mathrm{E}[\log(C/H)\,|\,H=h]$, where $h>0$. The regression tree is then enhanced by bagging once again. The explanatory variables made available to the tree model fitting procedure include all previously used predictors and the estimate of the conditional probability that there is non-zero catch in a cell. The bagging estimate for cell *i* is therefore:

$$\mathrm{E}\left[\log(C_i/H_i)\,|\,(H_i=h_i>0)\cap C_i>0\right]\cong\tfrac{1}{M}\sum_{j=1}^{M}\mathrm{BSE}_j\left[\log(C_i/H_i)\,|\,(H_i=h_i>0)\cap C_i>0\right] \tag{3}$$

The second stage in the process is to estimate *CPUE* in a cell given hooks. Provided $\mathrm{P}(H_i>0)>0$, this can be done using:

$$CPUE_i\cong\mathrm{P}(C_i\,|\,H_i=h_i>0)\exp\left\{\mathrm{E}\left[\log(C_i/H_i)\,|\,(H_i=h_i>0)\cap C_i>0\right]+\tfrac{1}{2}\sigma_{lc}^2\right\} \tag{4}$$

where the quantity σ_{lc}^2 used in the (naïve) bias correction for the back transformation of $\log(C/H)$ to $CPUE=C/H$ is estimated as:

$$\sigma_{lc}^2\cong\tfrac{1}{n_c}\sum_{j=1}^{n_c}\left\{\log(c_j/h_j)-\mathrm{E}\left[\log(C_j/H_j)\,|\,(H_i=h_i>0)\cap C_i>0\right]\right\}^2$$

and n_c is the number of cells with non-zero catch.

In (4) the conditioning is at the actual number of hooks allocated for the fished cell. Another estimate of CPUE that applies for both fished and unfished cells is the analogous quantity evaluated at the expected number of hooks:

$$CPUE_i\cong\frac{\mathrm{E}[C\,|\,H=\mathrm{E}[H_i]]}{\mathrm{E}[H_i]} \tag{5}$$

which is well-defined provided $\mathrm{E}[H_i]>0$, or at least this property holds for the estimate of $\mathrm{E}[H_i]$, which, as we have noted, is effectively guaranteed by the definition of the fishery. Equation (5) will be used as the main component of the annual *CPUE* index based on all cells in the fishery.

Statistical software

The ***R*** statistical environment (see http://cran.r-project.org) was used to run the analyses in this paper. The `ipred` package for ***R*** was used to perform the bagging analyses, which in turn needed `rpart`, `MASS`, `mlbench`, `survival`, `class` and `nnet` packages for ***R*** also to be available.

Results

Figure 1 shows the bagged estimates for the yearly *CPUE* of all the SBT caught. That is, the results are based on the total catch (in numbers of fish) and not a proportion of the catch representing a particular age group. The yearly estimates are calculated for all cells that have appeared at least once in the 32 years of data. The figure gives the estimated *CPUE* for the fished areas in any particular year, the unfished areas in the same year and the combined fished and unfished areas. The annual *CPUE* estimates "Fished cells: E" are calculated using (5) and the "Fished cells: O" are calculated using (4). The estimates for the combined fished and unfished areas are based on (5). As a point of reference, Figure 1 also plots two empirical (or 'nominal') yearly *CPUE* series. The 'nominal: weighted' series is the yearly weighted arithmetic mean of the observed cells, and the 'nominal: catch / hooks' is the total catch for that year divided by the total hooks allocated to all cells.

One striking feature of the figure is how close the bagging estimates of the yearly *CPUE* for the fished areas are to the 'nominal: weighted' *CPUE*, suggesting that bagging is able to capture most of the complex spatial and temporal interactions evident in the observed data. A more interesting but just as striking a feature of the figure is the low estimated annual *CPUE* for the unfished cells for most years. Such a result implies that there is effective targeting and the primary reason no fishing took place in the unfished areas is that the catch would have been low.

Reliability checks

It is clearly critical to check that the results in Figure 1 are meaningful and reliable. One aspect of this is to estimate the prediction error or misclassification rate for each of the four components in the estimation process outlined in the methods section. For the first component – the estimation of the probability that a square is fished – bagging gives a misclassificat[illegible] rate of 13.6%. For the second component – the estimation of the conditional expected log hooks in a square, given there is effort – bagging gives a prediction error of 1.36. For the third component – the estimation of the conditional probability that there is a catch in the square, given the hooks in a square – bagging gives a misclassification rate of 7.3%. For the final component, – he estimation of the conditional expectation $\mathrm{E}[C/H\,|\,(H=h>0)\cap(C>0)]$ in a square – bagging gives a prediction error of 0.84.

Training and test data sets

A more comprehensive way to check the results in Figure 1 is to split the data randomly into two data sets: the training and test data sets that we have chosen to be equal sized. The training data are used to fit the model, including bagging, and the test data,[illegible]sed to validate the fitted model with comparable but independent data. This process may be repeated several times.

Running a full analysis on the training data and then fitting the model to the test SBT data gives a misclassification rate of 16.0% for the first component in the estimation process – the estimation of the probability that a square is fished – for the test data set. This result is not very different from the corresponding result in the previous paragraph in which the analyses were run with all the data included. The prediction error, defined as the average squared difference between observed and expected log hooks, is 2.37 for the second component of the estimation process – the estimation of the conditional expected log hooks in a square, given that there is effort – for that square in the test data. Such a value admittedly represents very high variability on the hooks scale, an effect we hope to improve if finer scale data and better non-spatial and non-temporal habitat predictors become available. This prediction error is slightly elevated relative to the corresponding result in which all the SBT data are included in the analysis. The misclassification rate is 8.8% for the third component of the estimation process – the estimation of the conditional probability that there is a catch in a square, given the hooks in a square – for the test data. This result is marginally higher than the corresponding result in which all the SBT data are included in the analysis. The prediction erro[illegible]0.84 for the final component of the estimation process – the estimation of $\mathrm{E}[C/H \mid (H=h>0)\cap(C>0)]$ for a square– for the test data. This prediction is the same as the corresponding result in which all the SBT data are included in the analysis.

The estimated yearly *CPUE* for the test data set can be found in Figure 2. The results show that for the test data set the estimated yearly *CPUE* for the fished areas is marginally lower than the empirical yearly *CPUE* of the test data, but captures all the important trends in the empirical yearly *CPUE*. In addition, the estimated yearly *CPUE* for the unfished areas in the test data is higher than the estimated yearly *CPUE* for the unfished areas in the whole data set (see Figure 1). Figure 2 in essence suggests that bagging is able to capture and model the main features in the SBT data.

To study further the usefulness of bagging an examination of the results for the individual cells in the test data set is useful. Figure 3 shows the fourth root of the observed number of hooks against the fourth root of the estimated number of hooks for the test data set. (Taking the fourth root is visually close to the log transformation but avoids the complication of zero values.) Figure 3 reveals that as the observed hooks in the test data set increase then in general the estimated hooks for the test data also increase. Figure 4 plots the same data plotted in Figure 3 but breaks it up by year. An examination of the plots in this figure suggests that the relationship between observed hooks and estimated hooks for the test data holds for all 32 years.

Figure 5 shows the fourth root of the observed *CPUE* plotted against the fourth root of the estimated *CPUE* for the test data set, calculated using (5). This figure illustrates that as the observed *CPUE* in the test data set increases then the estimated *CPUE* for the test data also increases. Only a very small number of moderate to large estimated *CPUE* values have observed *CPUE* values of zero. Figure 6 shows the same data as plotted in Figure 5 but breaks it up by year. This figure shows that the relationship between observed and estimated *CPUE* for the test data holds reasonably well for all 32 years, although the relationship does not appear to be as strong in the early years.

For comparison, Figure 7 shows the fourth root of the observed number of hooks against the fourth root of the estimated number of hooks for the training data set. As expected, it is clear from this figure that for the training data the agreement between the observed number of hooks and the estimated number of hooks is closer than that for the test data (see Figure 3). As for the test data, this figure illustrates that as the observed number of hooks in the training data set increases then the estimated number of hooks for the training data set also increases. Similar comments can be made regarding Figure 8, which plots the fourth root of the observed *CPUE* values against the fourth root of the estimated *CPUE* values for the training data set, calculated using (5).

To check how well bagging estimates the probability that there is effort in a square, it is useful to compare for the test data the bagging probabilities against bagging classification of the squares. Previously it was mentioned that the bagging probability of a square being fished is the mean of *M* replicate data sets. Another approach could be to classify the square according to that class which has majority in the *M* replicate data sets. Table 1 compares the results of such “majority vote” bagging with the average probability approach for the effort part of the modeling. The results in Table 1 show that when a square is classified as being unfished then most squares have bagging probability of being fished less than 0.38. When a square is classified as being fished then most squares have bagging probability of being fished greater than 0.56. This implies that the bagging probability approach has worked well.

Table 2 compares for the test data the results of “majority vote” conditional bagging classification for whether there is catch in a square with the conditional bagging probability that there is catch in the square given the log hooks in the square. The results in Table 2 show that when a square is conditionally classified as having no catch then most squares have conditional bagging probability of there being catch less than 0.49. When a square is conditionally classified as having catch then most squares have conditional bagging probability of there being catch greater than 0.66. Once again, this implies that the bagging probability approach has worked well.

Checks on the methodology using simulation studies

Strongly empirical models such as the one we have adopted here could be seen as liable to detect spurious patterns in the data due to an implicit over-fitting. Another question that naturally arises is the extent to which this kind of modelling manages to capture and make use of spatial information. To investigate these issues in the present case we propose to run the same fitting process to data sets which are progressively more degraded by internal randomisation. Four such simulated data sets were used. All involved the permutation of existing data. The permutation of the data was spatial in nature only in the sense that for any datum the month component remains correct but the spatial components may change. Nevertheless even if the month component is correct, varying the cell fished to a different area within a month in effect injects some temporal change.

The four schemes used were:

1. The observed monthly effort values for cells were randomly permuted over cells simulating a new spatial distribution of effort and the observed *CPUE* values were randomly matched to the permuted effort values.
2. The observed monthly effort and *CPUE* pairs were randomly permuted over cells.
3. The *CPUE* values were kept fixed but the non-zero effort patterns were randomly permuted among each other. (In other words, those squares that had no observed effort in a month continued to have no effort, but the squares with positive observed effort kept the same *CPUE* but not necessarily the same effort.)
4. The existing pattern of effort was maintained but the *CPUE* values were randomly permuted amongst those squares with positive effort. (Similar to 3 with Effort and *CPUE* values interchanging roles.)

While maintaining some contact with the actual data, these schemes simulate different scenarios in an effort to tease out any weaknesses in the modelling strategy. The features of the real data that are randomly broken are the association of effort with *CPUE* (apart from zero effort, which is always associated with an unknown *CPUE* and vice versa), and the association of effort or CPUE with a specific spatial location. Temporal associations are not broken.

Note that the first two schemes alter the actual areas fished, that is cells that have no effort in the actual data can be given effort in these cases, whereas the last two schemes maintain the actual areas fished. In this sense the first two schemes may be considered to inject a greater degree of spatial disengagement with the real data.

As for the analyses of the non-permuted observed data, the simulated data were analysed using all the data and randomly splitting up the data into train and test data sets. Five simulations were run for each permutation scheme. Results from one simulation for each scheme are presented in large format for the test data sets and using all the data. The results for all five simulations for each scheme are presented in small format in Figure 17 to Figure 20, showing that the outcomes were typical.

Although viewing the model outcomes via a weighted annual *CPUE* may not be the most sensitive way of judging the internal workings of the procedure, it is the outcome of primary consequence and we limit our investigation to this consequential summary measure.

Figure 9 to Figure 12 show the estimated yearly *CPUE* values for representative simulations for the four schemes, respectively, when all the data are used; Figure 13 to Figure 16 show the same output diagrams for representative simulations for the four schemes for the test data sets, using a model developed from a complementary training set in the usual way.

An examination of Figure 9 and Figure 10 reveal that for the first two simulation schemes the enhanced tree-based method estimates more similar yearly *CPUE* values for the fished and unfished squares, than in the other two schemes. While the overall trend in annual *CPUE* will largely be dominated by the individual cell *CPUE*'s on which it is based, this result – the failure to unlink fished and unfished cells – is to be expected particularly in the first simulation scheme in which both the distribution of effort was random in space and the distribution of *CPUE* within cells was independent of effort (given that the effort was positive). These results suggest that spatial information is important and is being utilised by the model fitting process to separate *CPUE* in fished and unfished areas, although there has to be an element of speculation in this interpretation. Figure 11 and Figure 12 reveal that for the third and fourth simulation schemes the enhanced tree-based method distinguishes between fished and unfished cells, especially for the first half of this series, but not quite as successfully as for the original data. This suggests that even if the CPUE or effort figures are not correctly matched but the fished areas remain true, the effect on the annual estimate of CPUE is not all that much affected using this method. This strengthens our view that the spatial aspect is one important component in the prediction of the difference between the expected *CPUE* for fished and unfished squares. (Note that 'fished' and 'unfished' cells here mean according to the permuted data; in the latter two schemes these agree with the original data but in the former two they will not.) In addition, the fact the estimated differences between fished and unfished squares is substantially greater for the original data sets than in either of the latter two simulation schemes indicates that the joint and linked distribution of *CPUE* and effort is important. This suggests that an association of *CPUE* and effort (*i.e.* targeting) may be an additional important component in the estimation models prediction of the differences between the expected *CPUE* in fished and unfished squares.

We do not claim that the method is entirely free of over-fitting bias, however. Some insight on this, with care, can be got by looking at Figure 21, which presents three ratios

a) The ratio of CPUE estimates for fished cells using expected and observed hooks, respectively (blue line)

b) The ratio of CPUE estimates for fished cells using expected hooks to that for unfished cells (which can only used expected hooks – red line)

c) The ratio of CPUE estimates for fished cells using observed hooks to that for unfished cells (green line)

The first four panels correspond, in order, to the four simulation scenarios and the fifth panel shows the same information for the original data and model. Similar information is given in Figure 22 to Figure 25, which gives the same information for the simulation runs with the training and test data, in each case showing the same information for the original test data and model in the sixth position.

The first two panels in Figure 21 show that there is still some tendency for fished and unfished cells to separate, with fished cells having slightly higher expected CPUE. We can only explain this in terms of over-fitting bias, but comparing the situation with the final panel clearly shows that this bias is not enough to explain the effect detected in the real data. Even so, there appears to be an appreciable effect of over-fitting. The deviations in the ratio (fished / unfished) in the real data (final panel) are bigger than seen in the simulations where there should not be a difference. The third and fourth panels, where spatial information is preserved along with the correct CPUE or effort figure, respectively, show a more complex pattern where some relevant information – which has to be partially spatial – has clearly been retained. In the third panel it is not surprising that tossing away effort as predictor does not do much to this summary measure, with respect to fifth panel (real data). When the unfished CPUE estimate falls to a very low value, as happens for the fourth panel particularly, the ratio becomes somewhat unstable. The large ratios might be an artefact of the ratio measure; however, there is some time consistency in that there are large values of the ratio for several successive years. The information in the first four panels of Figure 21 are based on only one simulation run, so care needs to be taken when drawing inferences. No more simulations were run because of computing and time constraints. Many more simulations need to be run to see how often patterns and their magnitudes are repeated when there should be none.

Figure 22 to Figure 25, in our view, reinforce this interpretation of over-fitting bias, somewhat. In Figure 22 and Figure 23 the ratio of fished to unfished cell CPUE estimates lie much closer to unity and tend to oscillate around them, indicating that if an independent test data set is used the over-fitting effects diminish. However in the last two figures, which correspond to cases where spatial and some other pertinent information is preserved there is a pattern still apparent, although there is a large increase in variance. Notice that in Figure 24 and Figure 25 there is no very consistent pattern in the five simulations (in contrast with the results for the full data), again strengthening the suggestion of over-fitting effects emerging.

The separation between effort and *CPUE* is never quite complete. In a sense all the effort figures are known, with a lot of them for any year/month combination being zero. The *CPUE* figures are a mixture of positive, zero and unknown values, with all the unknown values always coupled with the zero efforts and vice-versa. The aim is to estimate these unknowns. Simulation schemes presented here do not entirely separate effort and *CPUE*. Even so, simulation schemes first, third and fourth set things up so effort provides no predictive power on *CPUE* when effort is positive, so it may be reasonable to conjecture that, for the real model, the zero effort has no predictive power on *CPUE*. For the second simulation scheme, however, positive effort does help predict *CPUE*, and thus presumably zero effort does too. Position does not affect the distribution of effort and *CPUE* in the second simulation scheme; hence prediction depends on the unconditional distribution of *CPUE*. The trouble with this scheme is that the distribution of

CPUE when effort is zero is unobserved, meaning that when estimating for none fished areas this is done using a biased sample of *CPUE* values.

Reliability check information for the simulation results presented in Figures 9 to 12 can be found in Table 3 and those for Figures 13 to 16 can be found in Table 4. The corresponding results from the actual data (full data set and test data set respectively) is also shown for comparison. An examination of the results in Table 3 reveals that when effort and *CPUE* are randomly permuted, either separately (first scheme) or together (second scheme), the misclassification error for the first component of the modelling (the estimation of the probability that a square is fished) increases nearly three fold. For these two schemes the prediction error for the second (the estimation of the conditional expected log hooks in a square, given there is effort) and fourth component (the estimation of $\mathrm{E}[C/H \mid (H=h>0) \cap (C>0)]$ in a square) of the modelling and the misclassification error for the third component (the estimation of the conditional probability that there is a catch in the square, given the hooks in a square) of the modelling also are markedly higher than is the case when the data are not permuted. For the third simulation scenario, where only effort is randomly permuted within the fished cells, only the prediction error of the second component of the modelling is markedly higher than in the non-permuted case. For the fourth simulation scenario, where only *CPUE* is randomly permuted within the fished cells, the misclassification error for the third component of the modelling and the prediction error for the fourth component of the modelling are markedly higher than in the non-permuted case. Similar remarks can be made for the reliability check information in Table 4. Note that in every case the third simulation scheme, where spatial and CPUE information is cogent but effort information is dissociated, these performance statistics are generally very comparable with the actual data case.

To ascertain what impact weighting has on the simulation outputs, Figure 26 to Figure 29 plot the non-weighted estimated yearly *CPUE* for the four simulation schemes when all the data are included. Figure 30 corresponds to Figure 21, except it plots for the four simulation schemes the ratio values for the non-weighted estimated yearly *CPUE*. An examination of Figure 21(d) and Figure 30(d) suggests that weighting actually reduces the extreme ratio values. The investigation of a variance-related weighting regime may be of some interest. Here, when effort is small, the *CPUE* will have greater variability.

Discussion

Shone *et al.* (2001) use the regression tree methods CHAID and CART to model the *CPUE* for the 1969 to 2000 SBT data. The two main components of the present approach, namely the four-stage conditional chain of models to capture both effort and catch in the fishery and the use of bagging to render stable the tree-model components and improve prediction accuracy make the present approach substantially different to that of Shone *et al.* (2001) and Watters and Deriso (2000).

Both key quantities, *H* and *C*, are modelled as mixtures of a saltus at zero, continued by a log-normal component for the positive values. This acknowledges that, in both cases, there is a non-negligible zero component but the positive component has approximately a constant coefficient of variation, requiring a log-transformation to be done prior to any least squares procedure that assumes constant variance. Regression trees are indeed estimated by least squares so transformations of the dependent variables to achieve homoscedasticity should always be considered.

Although, on their own, regression and classification trees are unstable as predictors, they are very effective for modelling discontinuous change and complex interactions between variables.

The individual trees are easily interpretable, even if this feature is lost through bagging. The analyses presented in the results section illustrate that the bagging of regression and classification trees can result in good estimates for expected number of hooks and the expected *CPUE* in a cell (see Figures 3, 5, 7 and 8). The good estimation of number of hooks, numbers caught and the expected *CPUE* in as cell holds for all 32 years of the data (see Figures 4 and 6). The estimation of the probability of a cell being fished and the conditional probability that cell has catch also is good (see Tables 1 and 2). The estimated yearly *CPUE* for the fished cells for the test data follows the empirical weighted yearly *CPUE* for the test data although it is marginally low for some years (see Figure 2). The estimated yearly *CPUE* for the fished cells for all the data is very close to the empirical weighted yearly *CPUE* for all the data (see Figure 1). The combination of bagging and regression and classification trees has led to results that accurately reflect large-scale features of the real data and have accounted for at least some of the discontinuous change and complex interactions between variables so evident in the SBT data.

Through simulations, we investigated the performance of the model under four scenarios which progressively removed the links between location, effort, and *CPUE*. The aim was to check whether the method would tend to find spurious patterns: in particular, when *CPUE* and effort were independently and randomly distributed in space, we would expect there to be no consistent difference between estimated *CPUE* in fished and unfished squares. The results of the simulations do suggest that the method may “discover” a certain amount of pattern where none exists, e.g. in terms of the average difference between estimated *CPUE* in fished and unfished squares. Nevertheless, the extent of pattern generated in the simulations through “over-fitting” was much less than that found in the real data, suggesting that the estimated differences between fished and unfished squares is not just an artefact of model-fitting. This conclusion is reinforced by the results from splitting the data into test and training sets. However, it is important to note that, for each of the four scenarios, time allowed only a single simulation replicate to be done. Until further replicates can be done, any general conclusions about the method’s performance (extent of over-fitting) or precision need to be made with caution.

Our results provide one plausible interpretation of the observed historical catch and effort data for SBT. However, we have not fully explored the ability of our models to cope with alternative hypotheses for why areas are sometimes not fished. This would require simulation-testing using mechanistic models of fish and fishery behaviour. Nevertheless, given the ability of the model to explain the data, it may be worth considering the CPUE indices from this method as one alternative measure of relative abundance in the context of SBT stock assessments and conditioning of operating models.

References

Breiman, L. (1996), Bagging predictors. *Machine Learning*, **24**, 124 – 140.

Breiman, L. (1998), Arcing classifiers (with discussion). *The Annals of Statistics*, **26**, 801 – 849.

Cressie, N. A. (1993), *Statistics for Spatial Data*. Revised edition. Wiley, NY.

Hearn, W. and Polacheck, T. (1998), 5 degrees standardised *CPUE*. CCBST/SC/9808/ 12.

Polacheck, T. and A. Preece. 2001. Exploratory and Descriptive Analyses of Catch, Effort and Size Data for Southern Bluefin Tuna from Japanese longline vessels. CCSBT/SC/0103/10

SAS (1999), *SAS/STAT User's Guide*, *Version 8*. SAS Institute Inc., NC.

Shono, H., Tsuji, S., Takahashi, N. and Itoh, T. (2001), Preliminary analysis of *CPUE* standardization and area stratification by tree-regression models. CCSBT/SC/0108 /30.

Thomas, M. and Toscas, P.J. (1997), Spatial analysis of Southern Bluefin Tuna catch per unit effort data for 1969 – 1996: A best linear unbiased predictor approach. CCSBT/SC/97/22.

Toscas, P. J., Thomas, M and Polacheck, T. (1998), Spatial analysis of Southern Bluefin Tuna catch per unit effort for 1969 – 1997: A best linear unbiased predictor approach. CCSBT/SC/9808/10.

Toscas, P. J., Venables, W. N., and Polacheck, T. (2001), Spatio-temporal analysis of Southern Bluefin Tuna catch per unit effort data: A best linear unbiased predictor approach. CCSBT/-SC/0108/??.

Watters and Deriso (2000), Catch per unit effort of Bigeye Tuna: a new analysis with regression trees and simulated annealing. Inter-American Tropical Tuna Commission, Bulletin **21**.

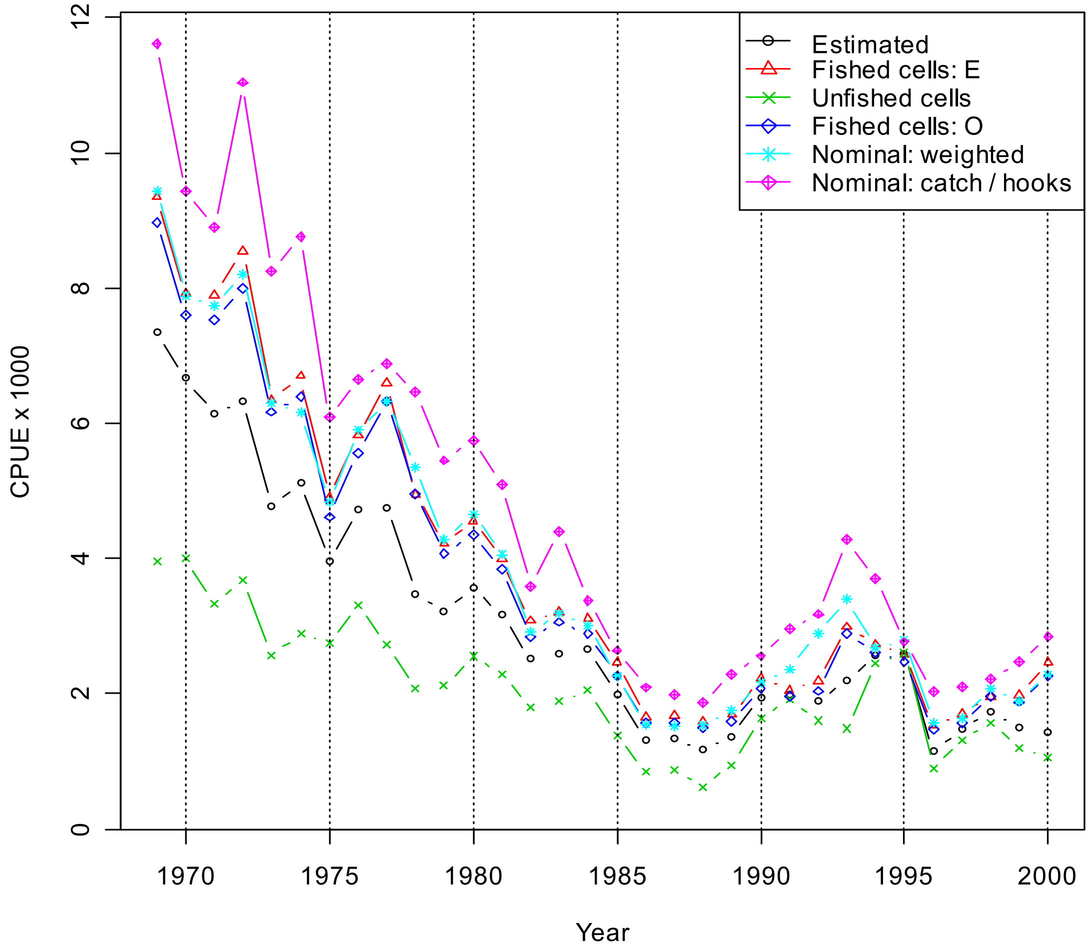

Figure 1: Nominal and tree-based estimates of CPUE, with separate contributions from fished and unfished cells.

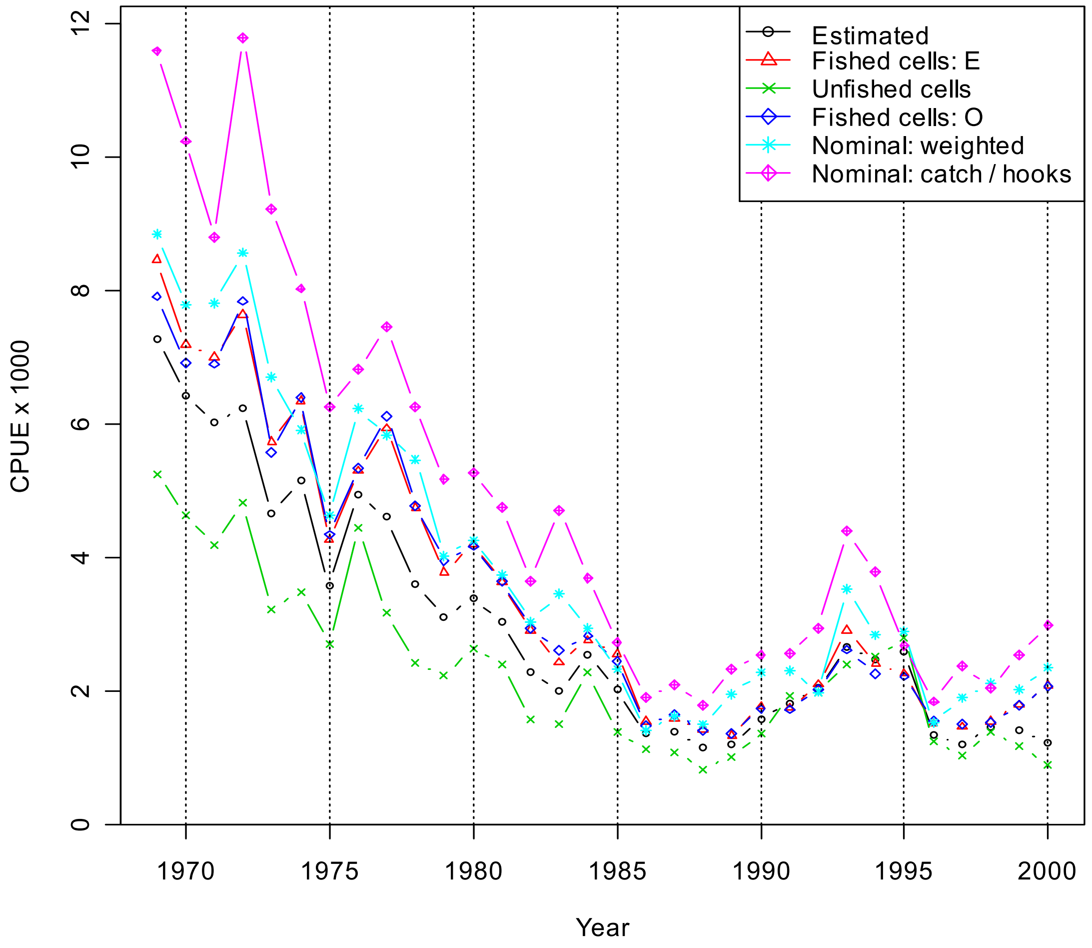

Figure 2: The empirical and estimated bagging yearly CPUE for the fished and unfished cells in the test data set.

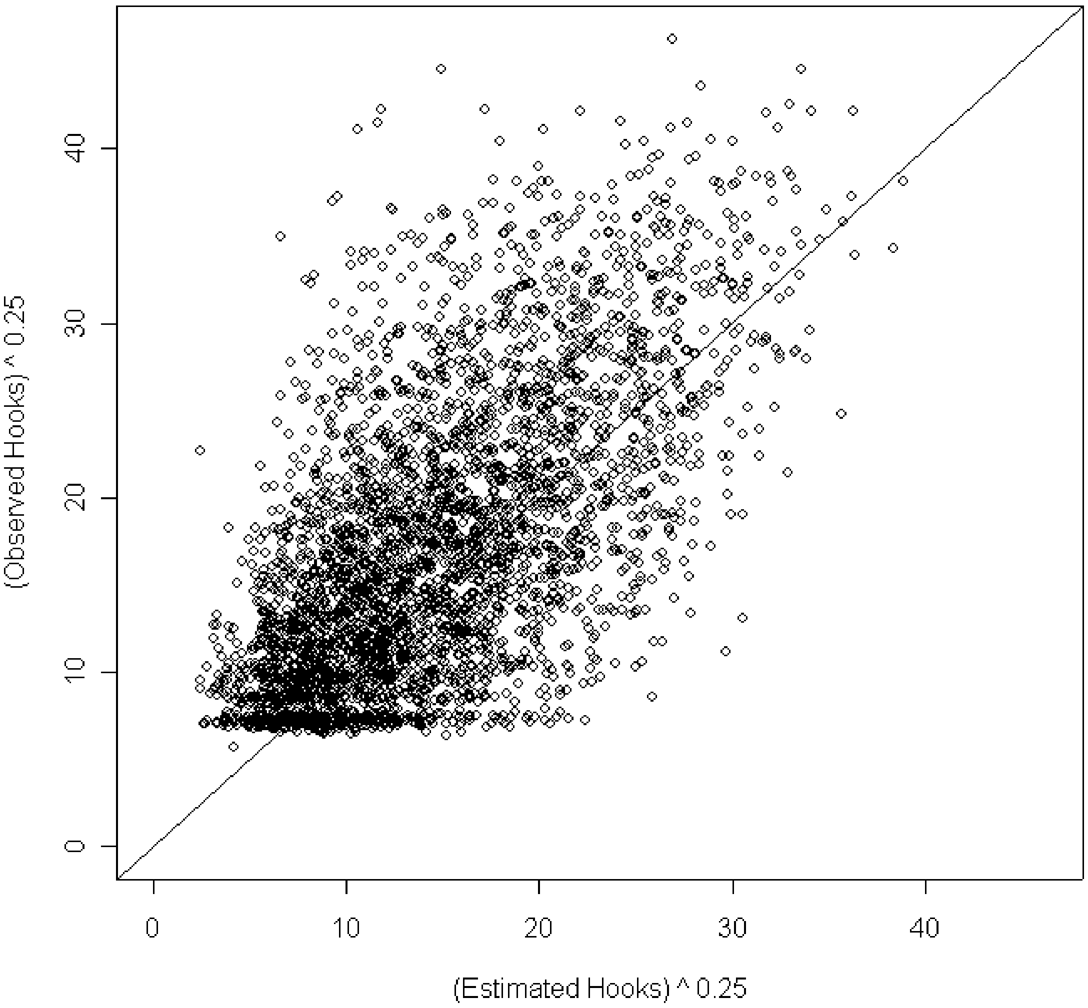

Figure 3: Observed vs estimated number of hooks, both fourth-root transformed

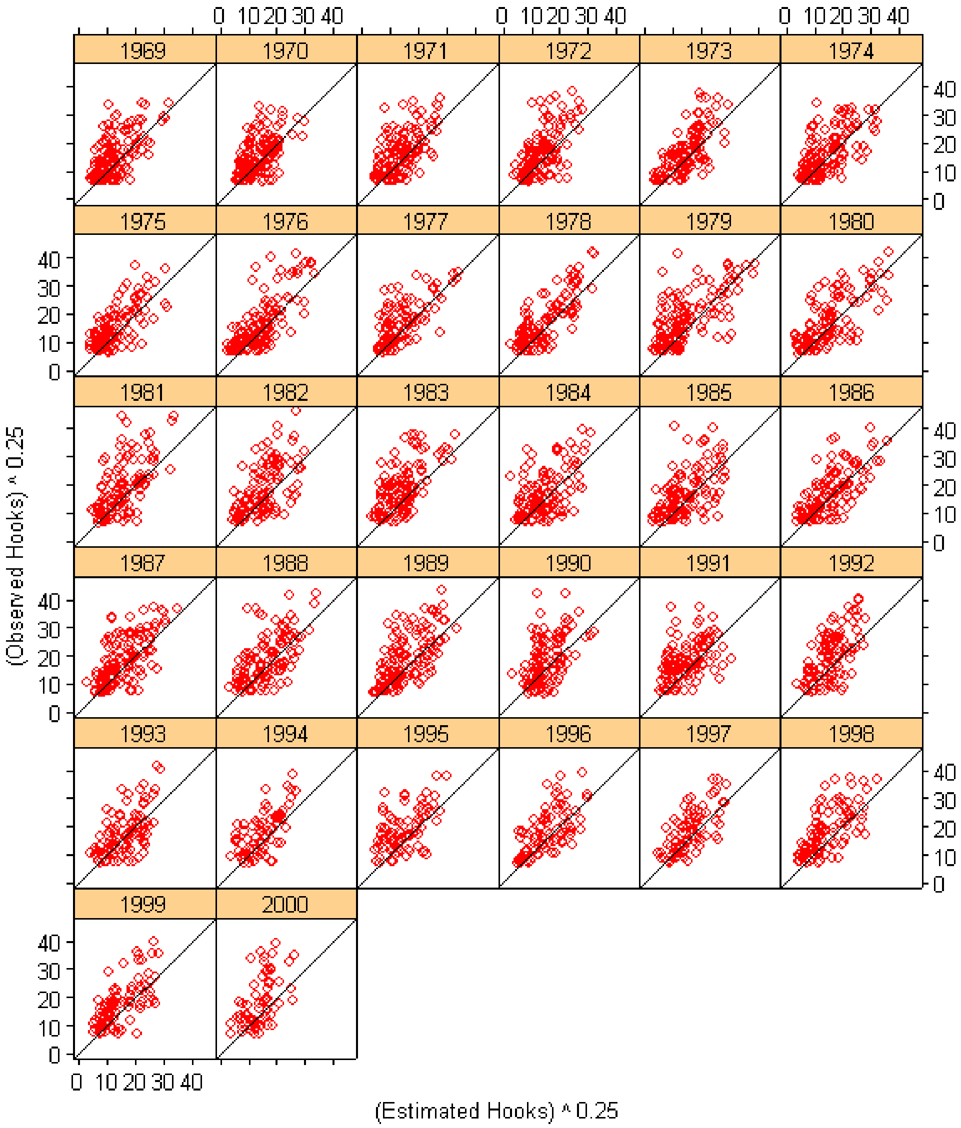

Figure 4: Yearly plots of the fourth root of observed number of hooks against the fourth root of estimated number of hooks for the test data.

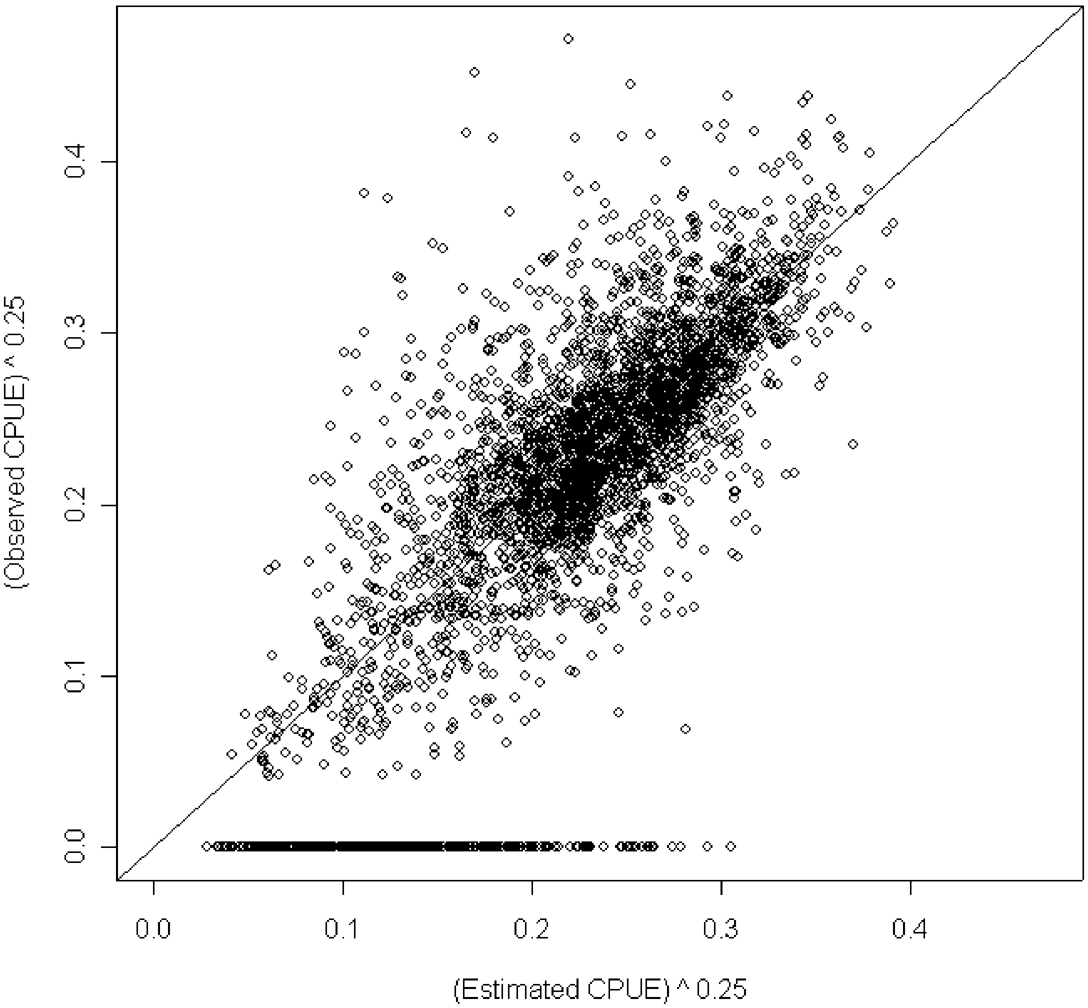

Figure 5: The fourth root of observed CPUE against the fourth root of estimated CPUE for the test data.

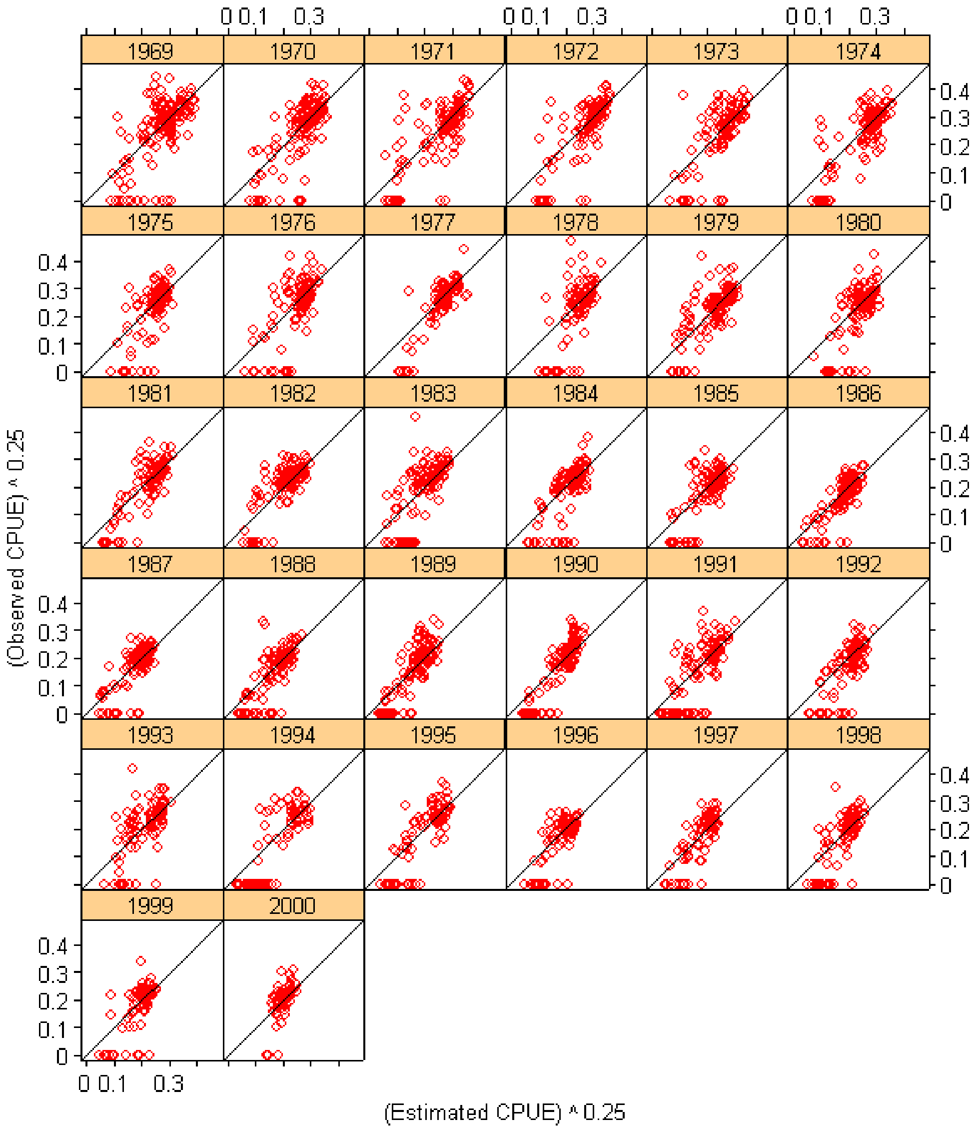

Figure 6: Yearly plots of the fourth root of observed CPUE against the fourth root of estimated CPUE for the test data.

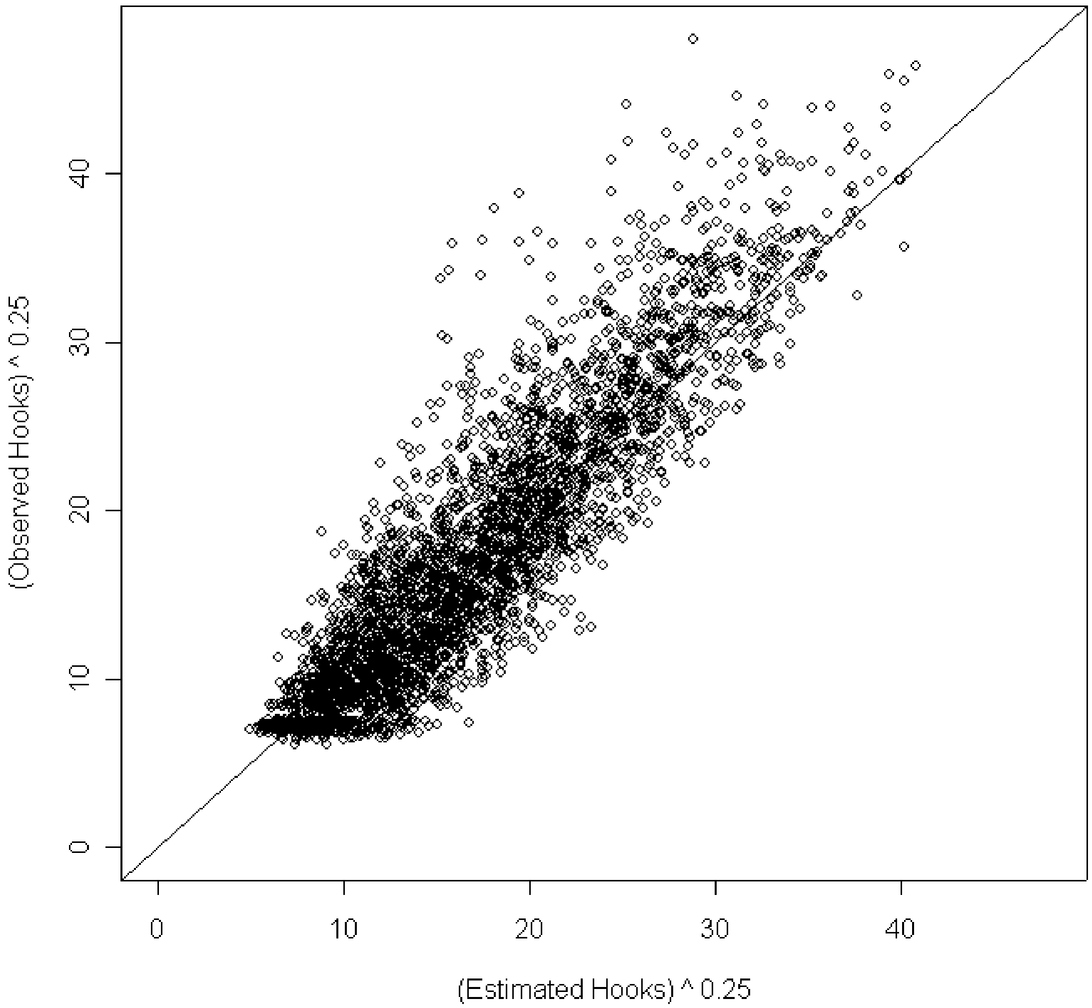

Figure 7: Plot of the fourth root of observed number of hooks against the fourth root of estimated number of hooks for the training data.

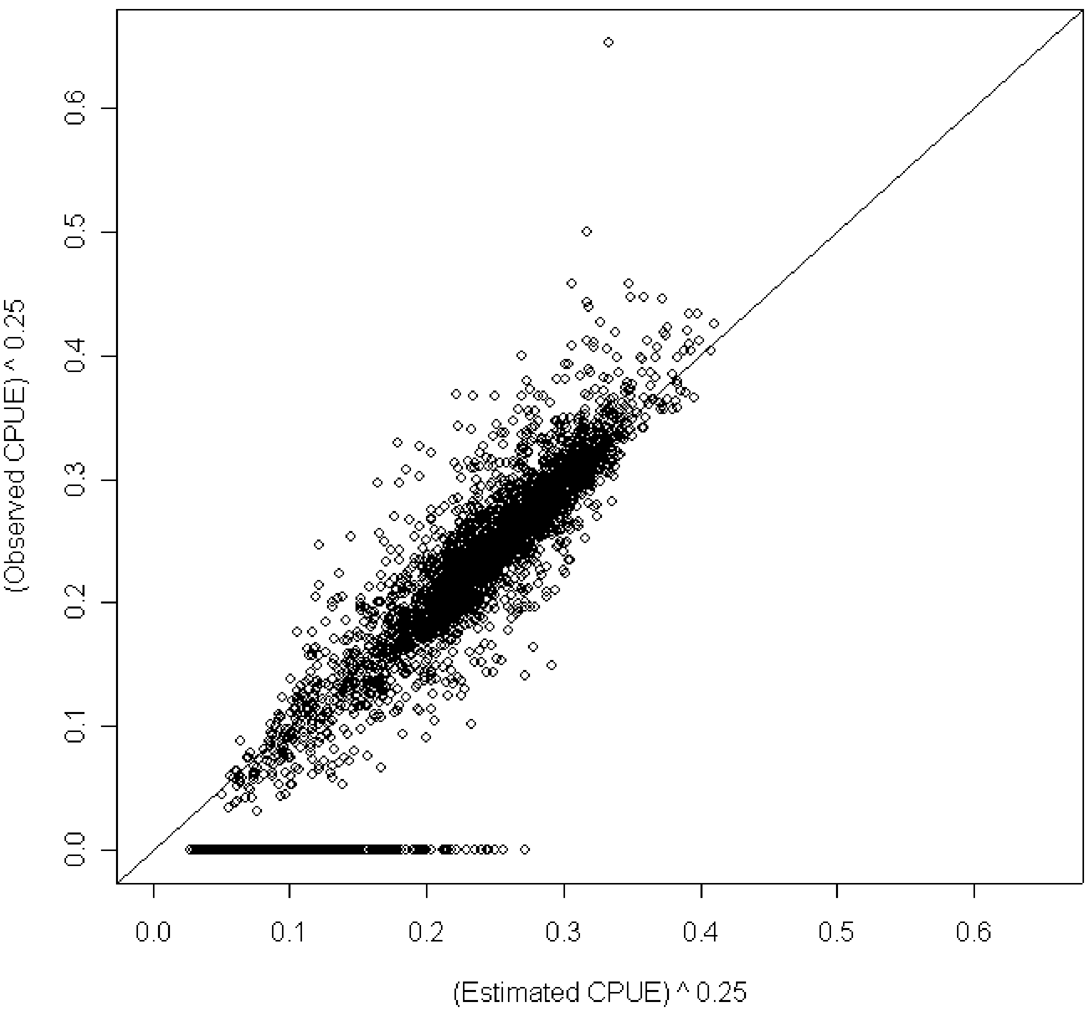

Figure 8: Plot of the fourth root of observed catch against the fourth root of estimated catch for the training data.

Table 1: This table presents for the test data percentiles for the bagging probability of there being effort in squares with bagging classifications as not having and having effort.

Percentile	Bagging probability for squares classified as having no effort	Bagging probability for squares classified as having effort
0	0.0048	0.4448
10	0.0134	0.5604
20	0.0286	0.6238
30	0.0514	0.6894
40	0.0808	0.7421
50	0.1193	0.8035
60	0.1696	0.8606
70	0.2245	0.9076
80	0.2960	0.9429
90	0.3791	0.9730
100	0.5537	0.9955

Table 2: This table presents for the test data percentiles for the conditional bagging probability of there being catch in squares with conditional bagging classifications as not having and having catch.

Percentile	Conditional Bagging probability for squares classified as having no catch	Conditional Bagging probability for squares classified as having catch
0	0.0161	0.4853
10	0.1267	0.6617
20	0.1936	0.7712
30	0.2423	0.8612
40	0.2823	0.9182
50	0.3249	0.9762
60	0.3610	0.9942
70	0.4079	0.9950
80	0.4520	0.9981
90	0.4886	0.9994
100	0.5656	0.9994

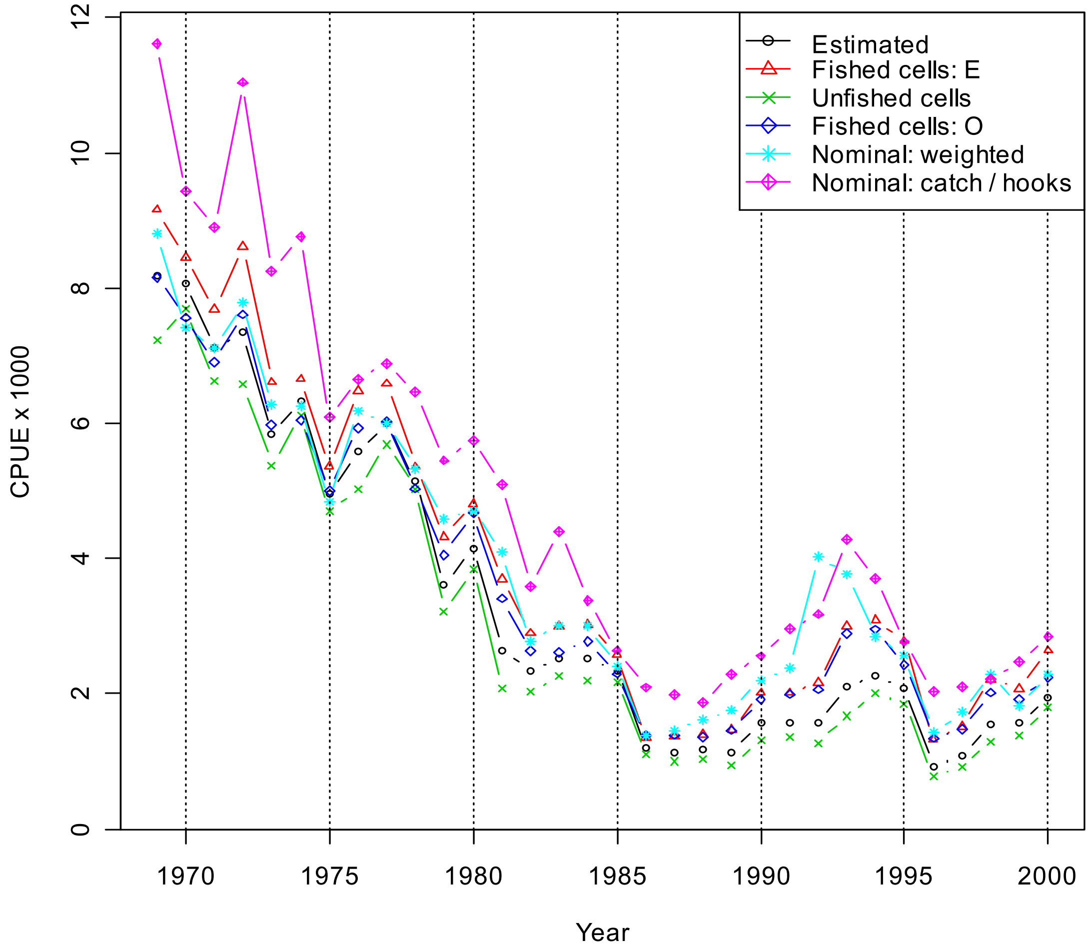

Figure 9: The empirical and estimated bagging yearly CPUE for the fished and unfished cells when effort is randomly permuted in space and CPUE is then randomly permuted in space to those cells with effort.

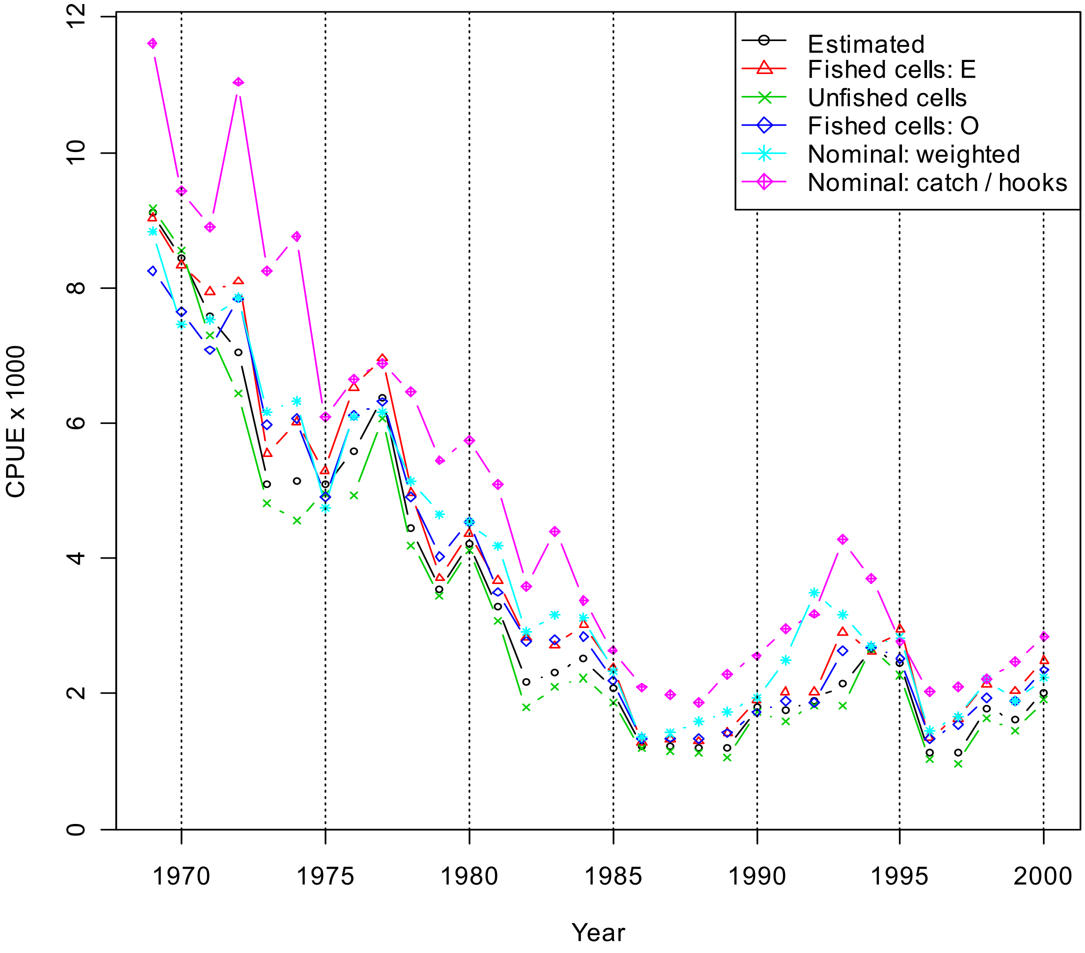

Figure 10: The empirical and estimated bagging yearly CPUE for the fished and unfished cells when effort and CPUE are randomly permuted in space.

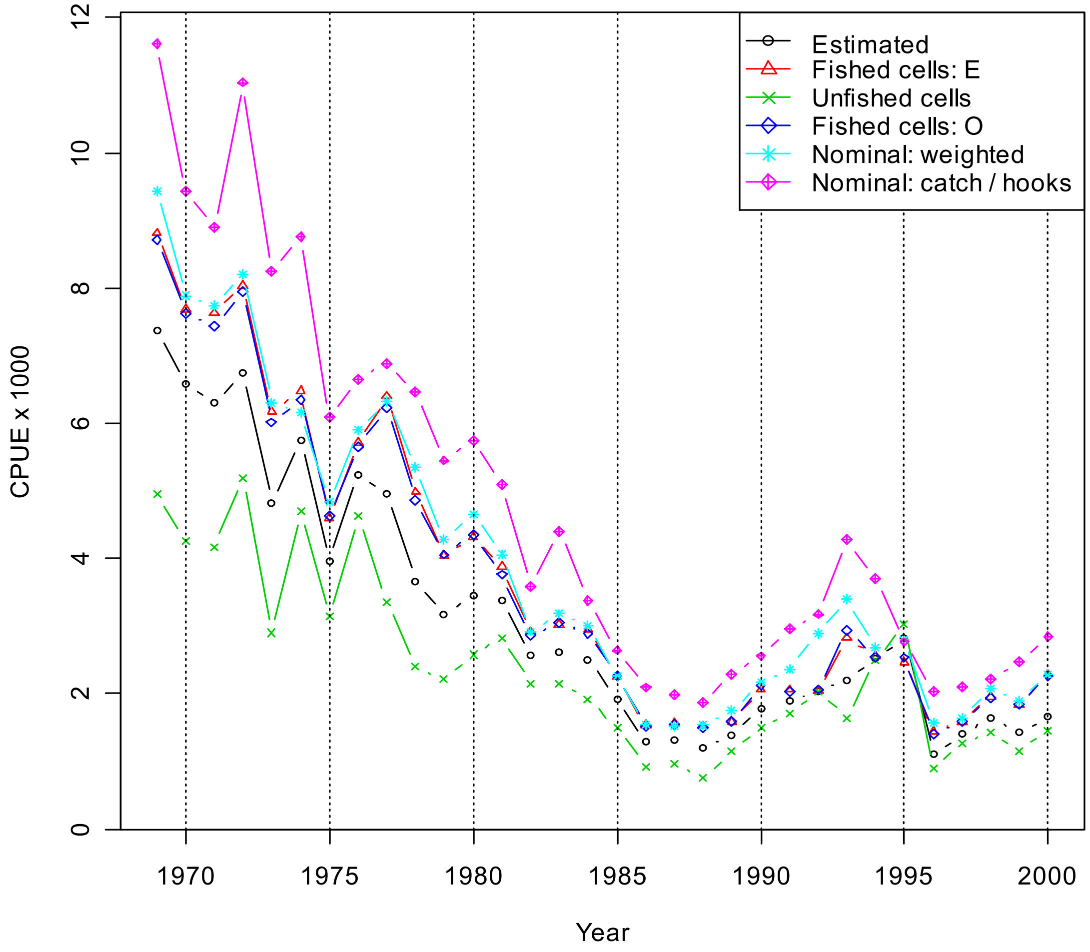

Figure 11: The empirical and estimated bagging yearly CPUE for the fished and unfished cells when the spatial pattern for CPUE is unchanged, but effort is randomly permuted in space for those cells where there is effort.

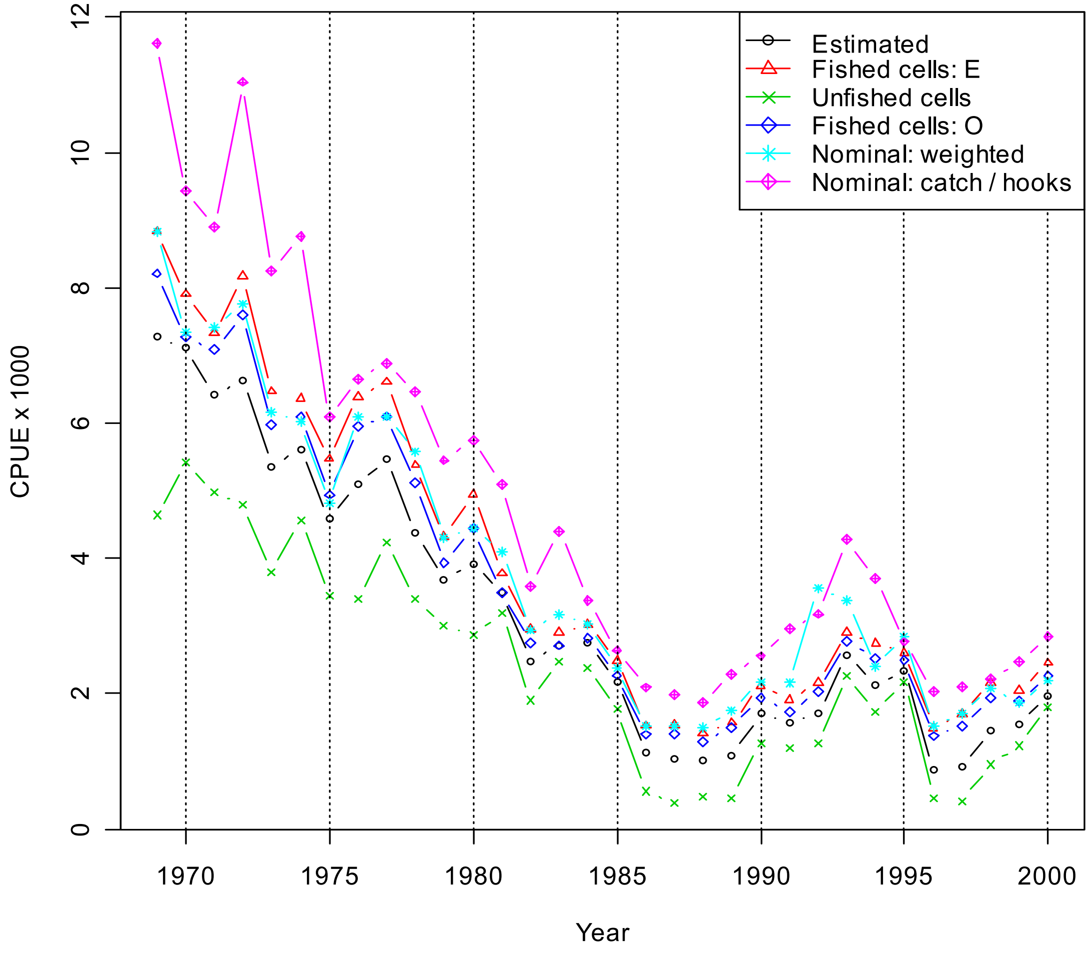

Figure 12: The empirical and estimated bagging yearly CPUE for the fished and unfished cells when the spatial pattern of effort is unchanged, but CPUE is randomly permuted in space for those cells where there is positive effort.

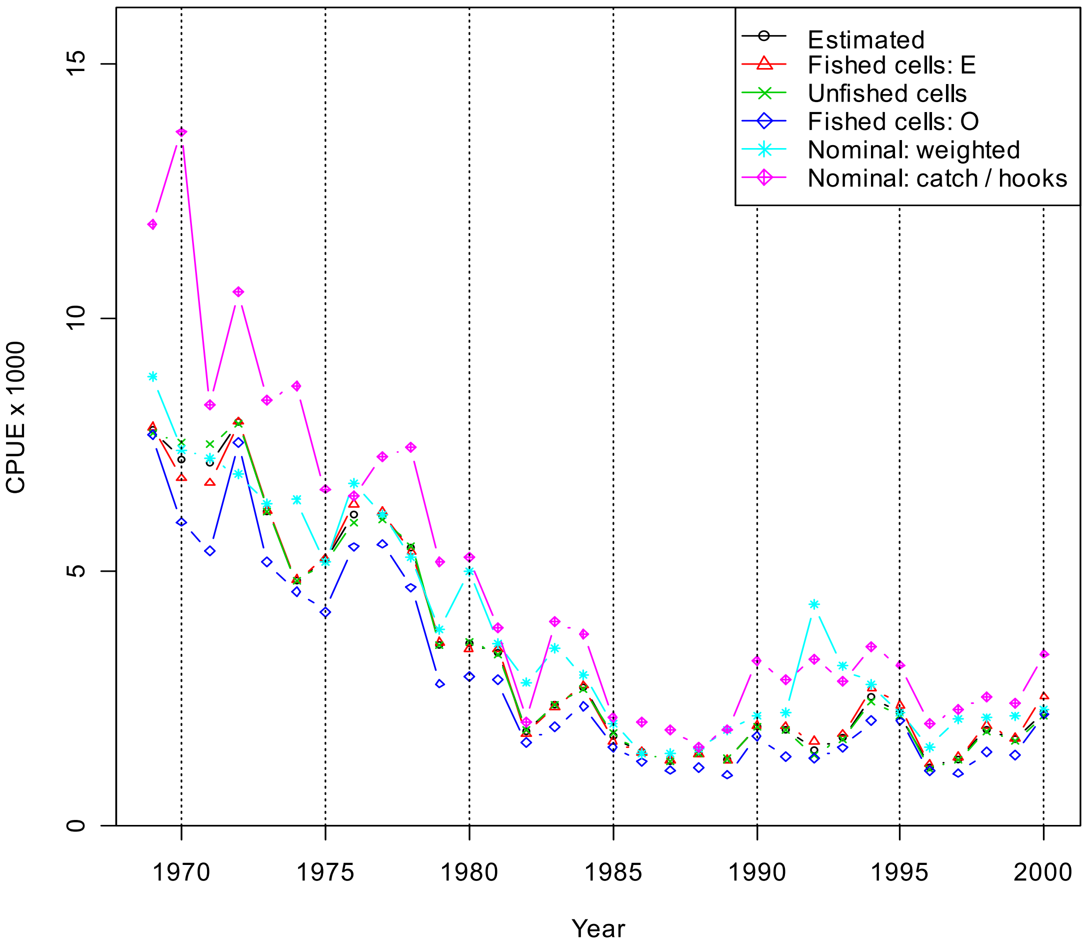

Figure 13: The empirical and estimated bagging yearly CPUE for the fished and unfished cells in the test data set when effort is randomly permuted in space and CPUE is then randomly permuted in space to those cells with effort.

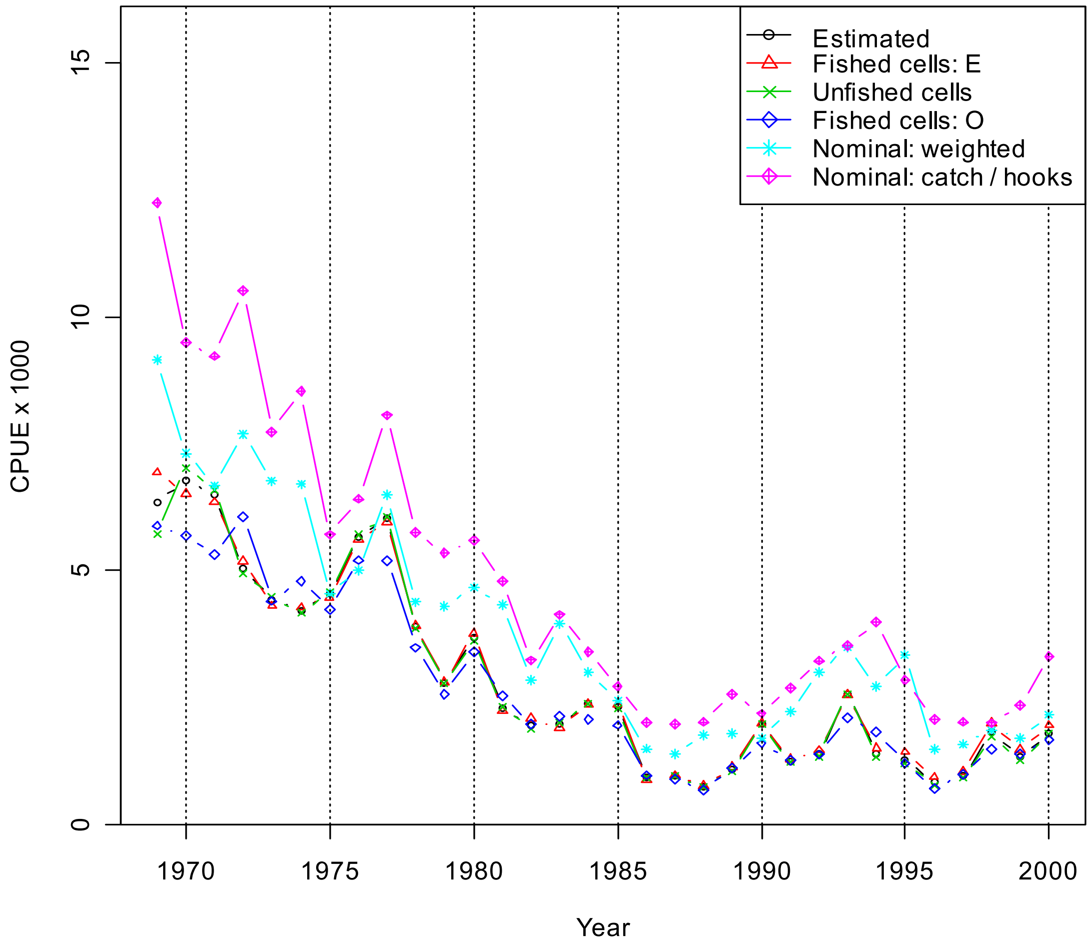

Figure 14: The empirical and estimated bagging yearly CPUE for the fished and unfished cells in the test data set when effort and CPUE are randomly permuted in space.

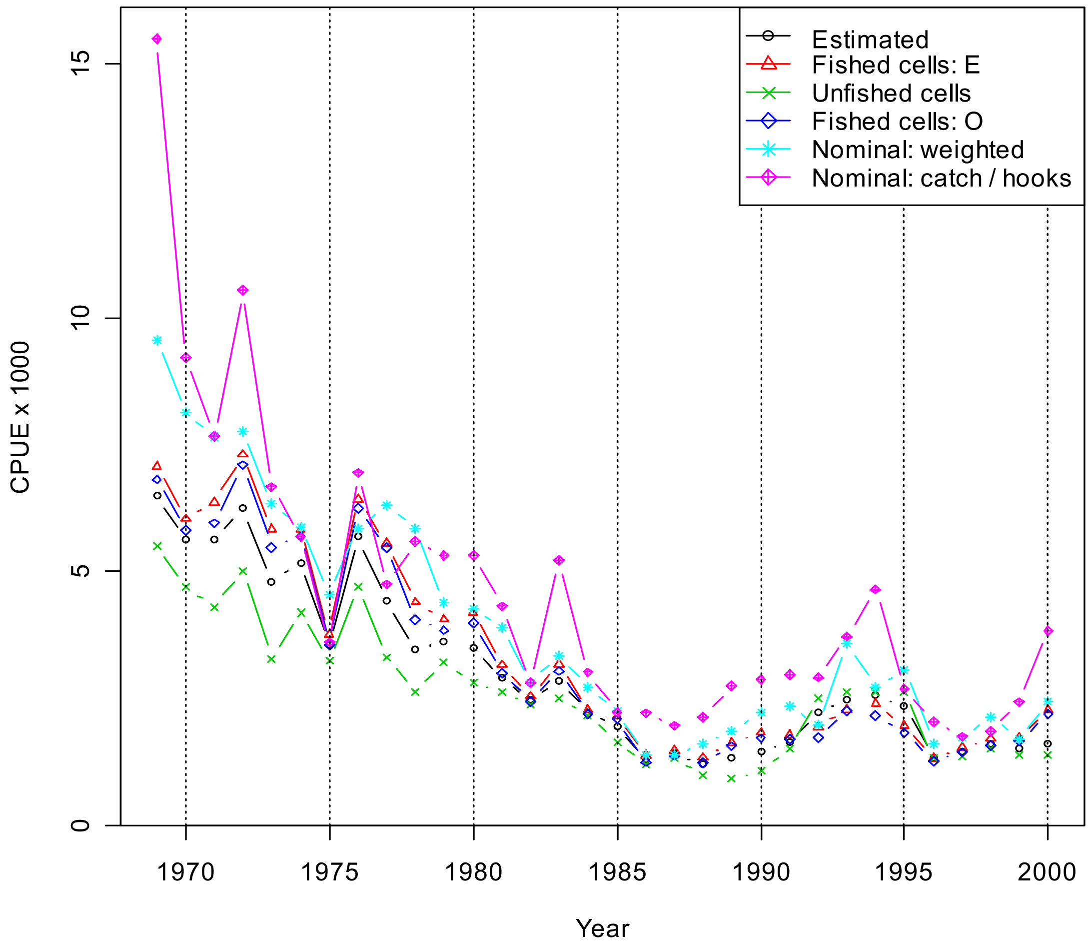

Figure 15: The empirical and estimated bagging yearly CPUE for the fished and unfished cells in the test data set when the spatial pattern for CPUE is unchanged, but effort is randomly permuted in space for those cells where there is effort.

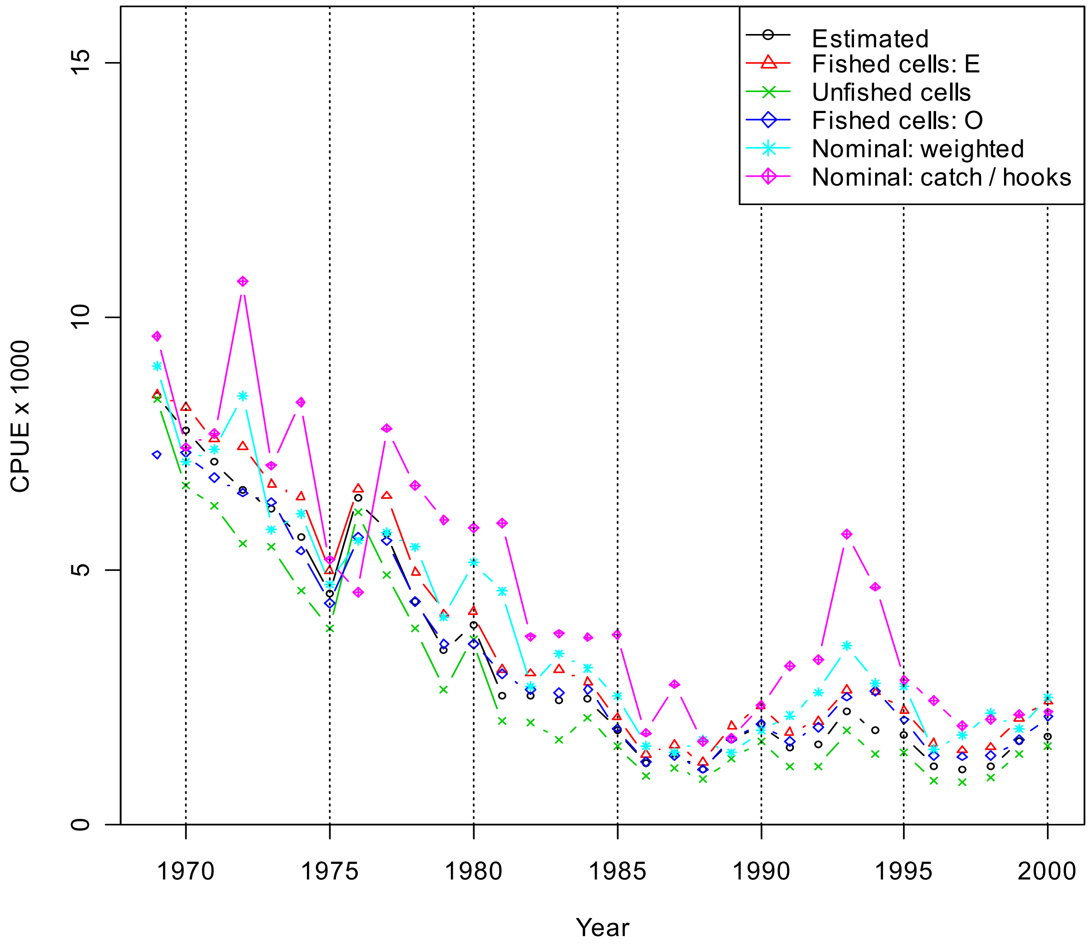

Figure 16: The empirical and estimated bagging yearly CPUE for the fished and unfished cells in the test data set when the spatial pattern of effort is unchanged, but CPUE is randomly permuted in space for those cells where there is effort.

Table 3: Reliability check results for the original observed data and the four simulation schemes shown in Figure 9 to Figure 12.

			Simulation scheme			
Modelling component	**Reliability check**	**Observed data**	**First**	**Second**	**Third**	**Fourth**
First	Misclassification error	13.6%	39.6%	39.4%	13.6%	13.4%
Second	Prediction error	1.36	1.98	1.97	1.96	1.34
Third	Misclassification error	7.3%	13.5%	14.1%	7.8%	13.7%
Fourth	Prediction error	0.84	1.29	1.27	0.86	1.29

Table 4: Reliability check results for the test data with the real data model, and the four simulation schemes for the test data sets depicted in Figure 13 to Figure 16.

			Simulation scheme			
Modelling component	**Reliability check**	**Test data**	**First**	**Second**	**Third**	**Fourth**
First	Misclassification error	16.0%	38.6%	38.8%	15.3%	15.6%
Second	Prediction error	2.37	3.82	3.75	3.88	2.36
Third	Misclassification error	8.8%	14.1%	15.9%	8.6%	13.7%
Fourth	Prediction error	0.84	1.69	1.74	0.86	1.61

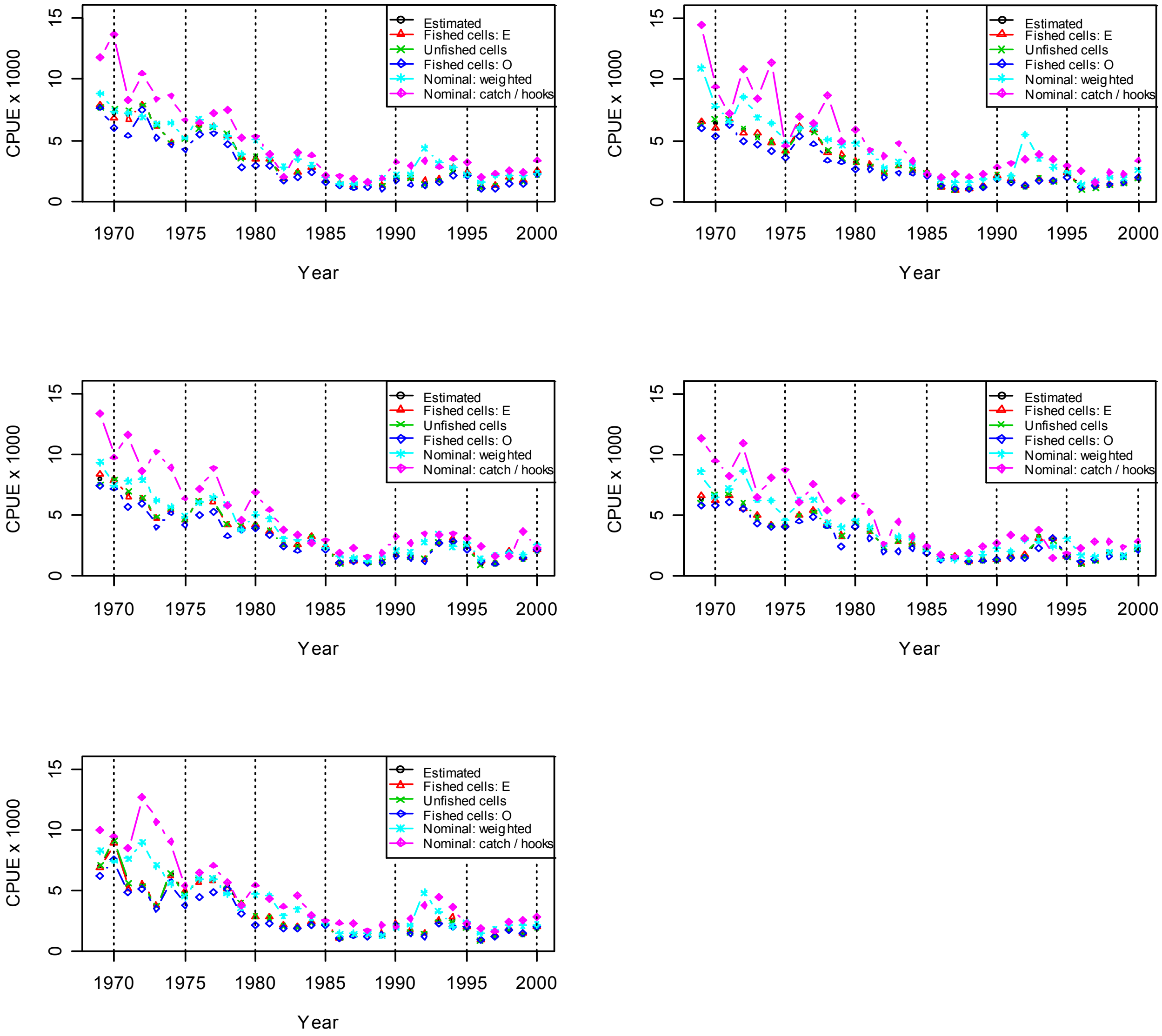

Figure 17: The empirical and estimated bagging yearly CPUE for the fished and unfished cells of five simulated test data sets when effort is randomly permuted in space and CPUE is then randomly permuted in space to those cells with effort.

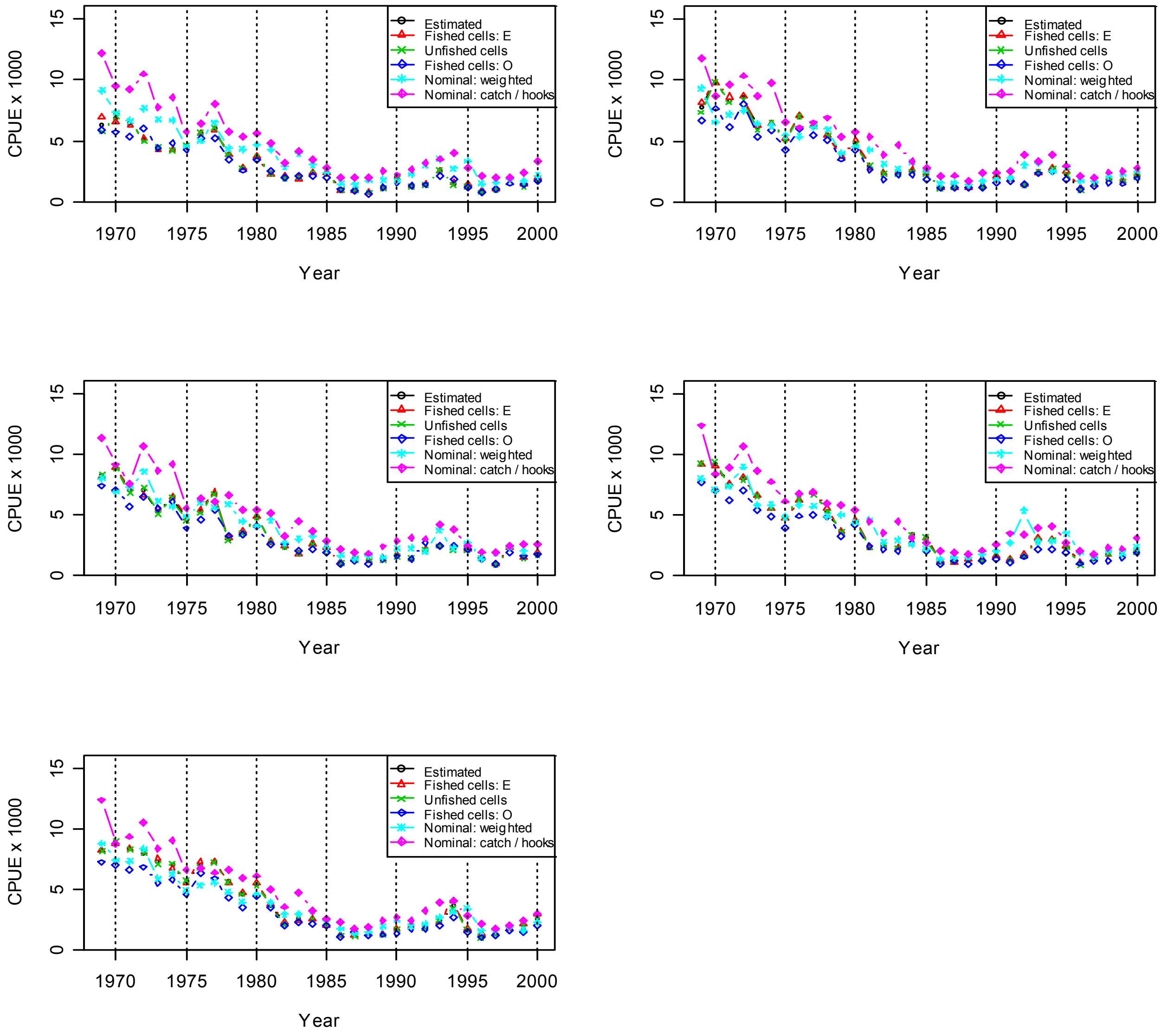

Figure 18: The empirical and estimated bagging yearly CPUE for the fished and unfished cells for five simulated test data sets when effort and CPUE are randomly permuted in space.

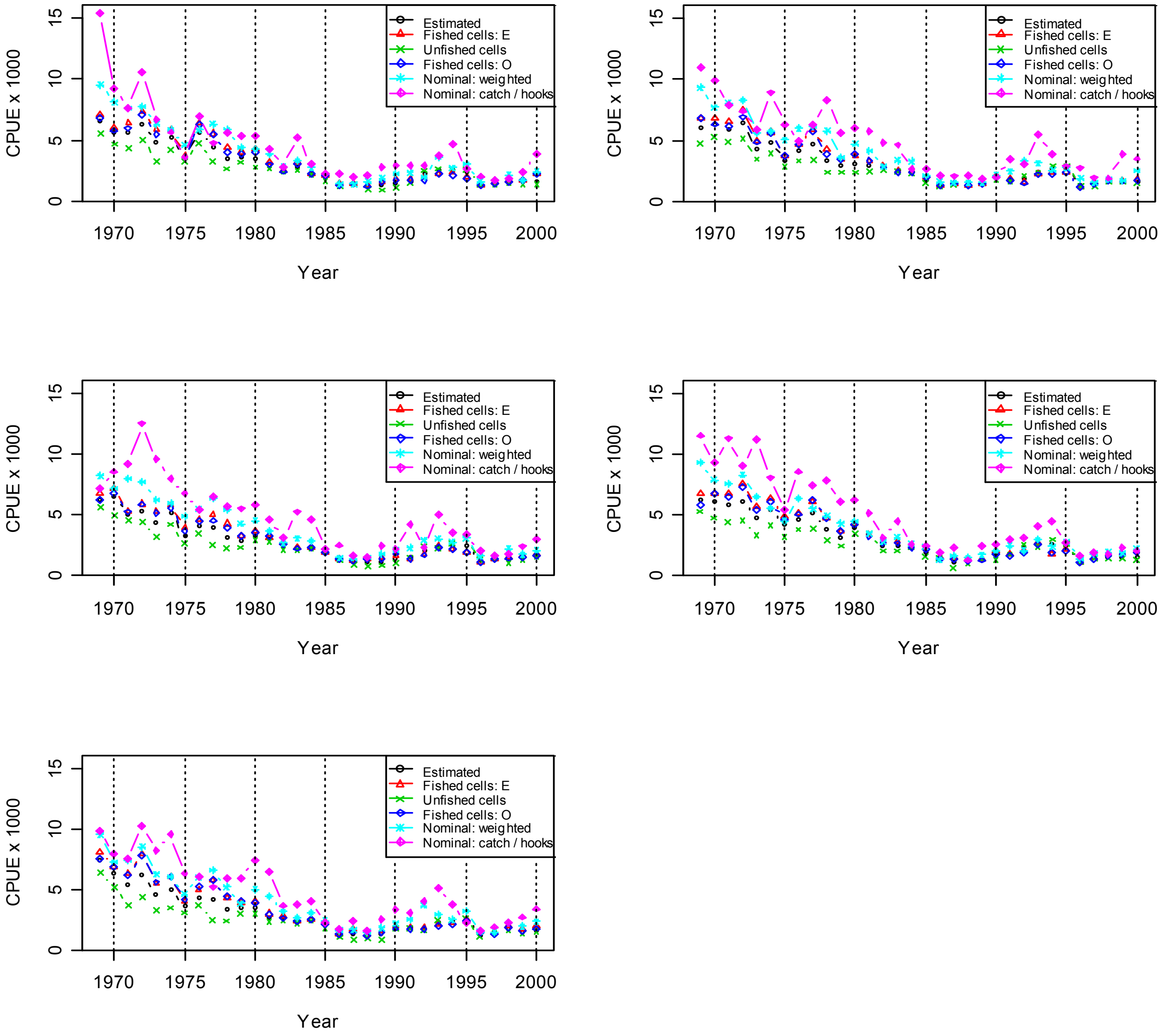

Figure 19: The empirical and estimated bagging yearly CPUE for the fished and unfished cells of five simulated test data sets when the spatial pattern for CPUE is unchanged, but effort is randomly permuted in space for those cells where there is effort.

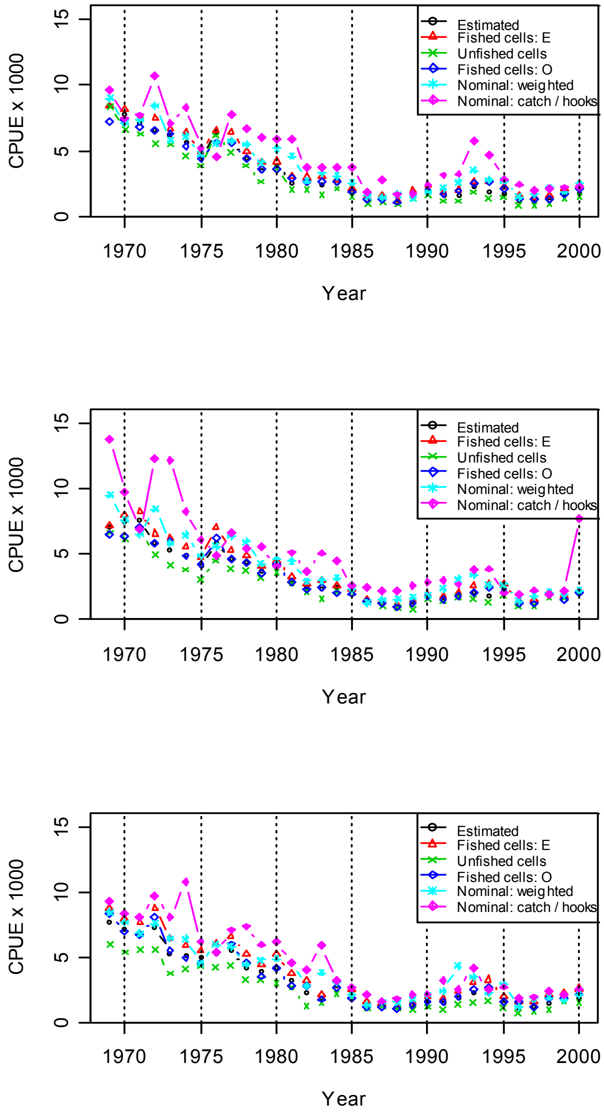

Figure 20: The empirical and estimated bagging yearly CPUE for the fished and unfished cells of five simulated test data sets when the spatial pattern of effort is unchanged, but CPUE is randomly permuted in space for those cells where there is effort.

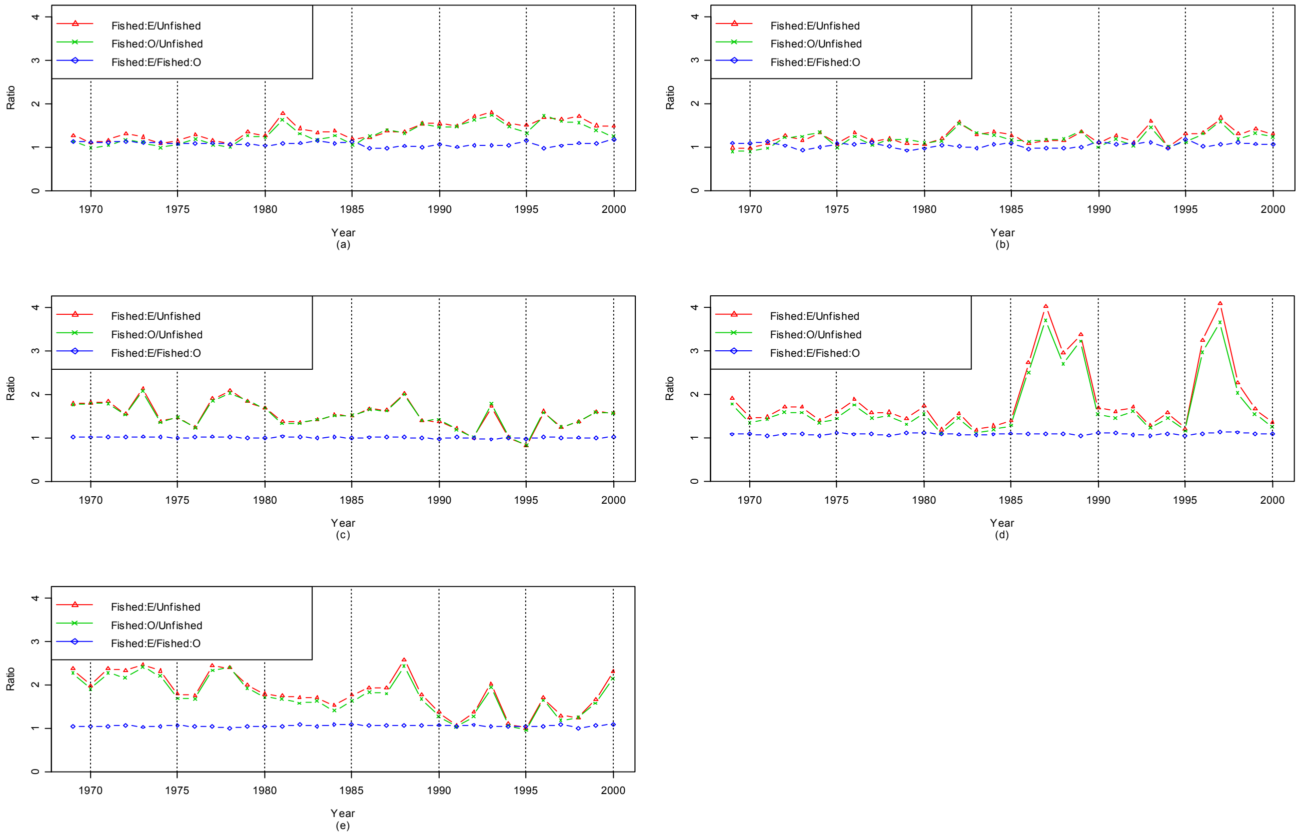

Figure 21: The ratio of CPUE for fished and unfished cells for (a) when effort is randomly permuted in space and CPUE is then randomly permuted in space to those cells with effort; (b) when effort and CPUE are randomly permuted in space; (c) when the spatial pattern of CPUE is unchanged, but effort is randomly permuted in space for those cells where there is effort; (d) when the spatial pattern of effort is unchanged, but CPUE is randomly permuted in space for those cells where there is positive effort; and (e) original observed data.

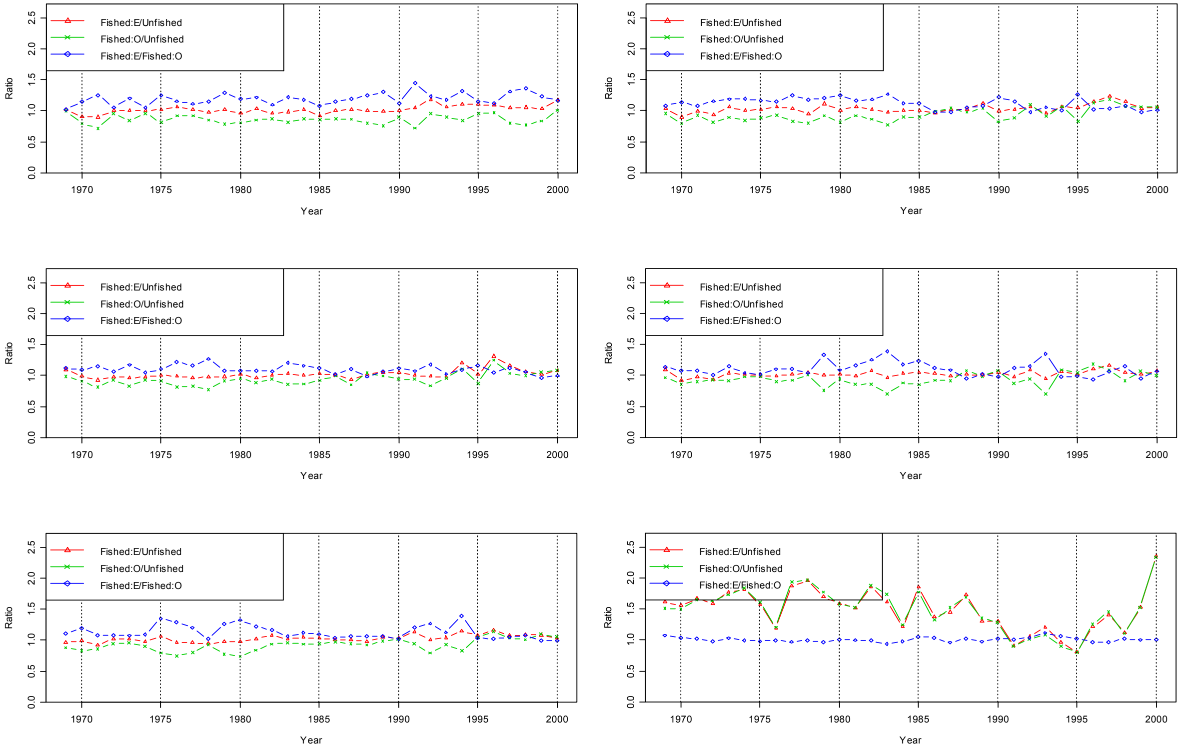

Figure 22: The ratio of CPUE for fished and unfished cells of five simulated test data sets when effort is randomly permuted in space and CPUE is then randomly permuted in space to those cells with effort. The last panel shows the same information for the original model and test data.

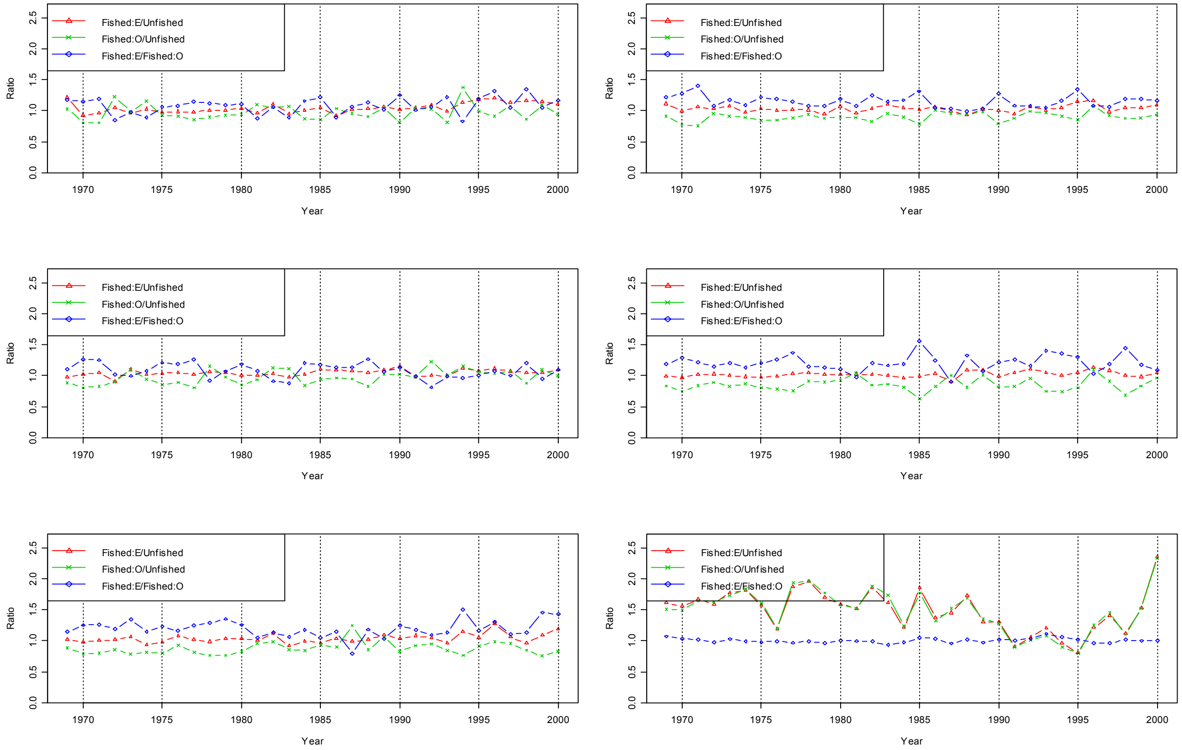

Figure 23: The ratio of CPUE for fished and unfished cells of five simulated test data sets when effort and CPUE are randomly permuted in space. The last panel shows the same information for the original model and test data.

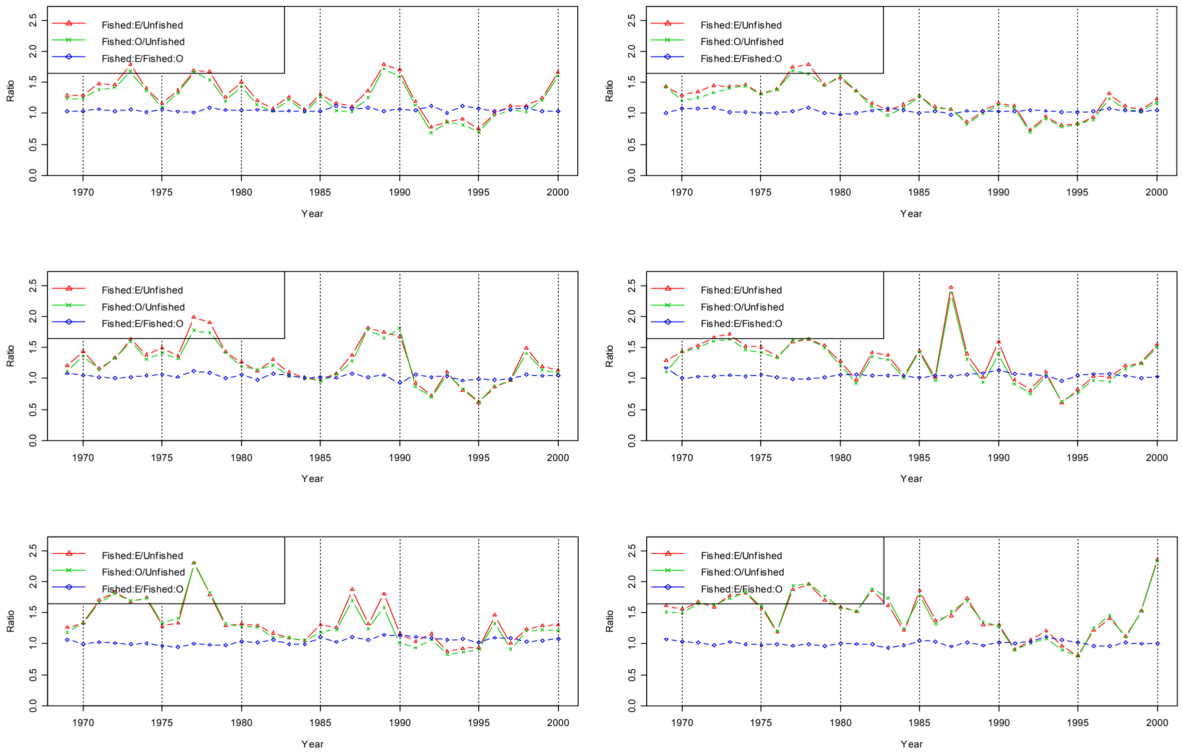

Figure 24: The ratio of CPUE for fished and unfished cells of five simulated test data sets when the spatial pattern of CPUE is unchanged, but effort is randomly permuted in space for those cells where there is effort. The last panel shows the same information for the original model and test data.

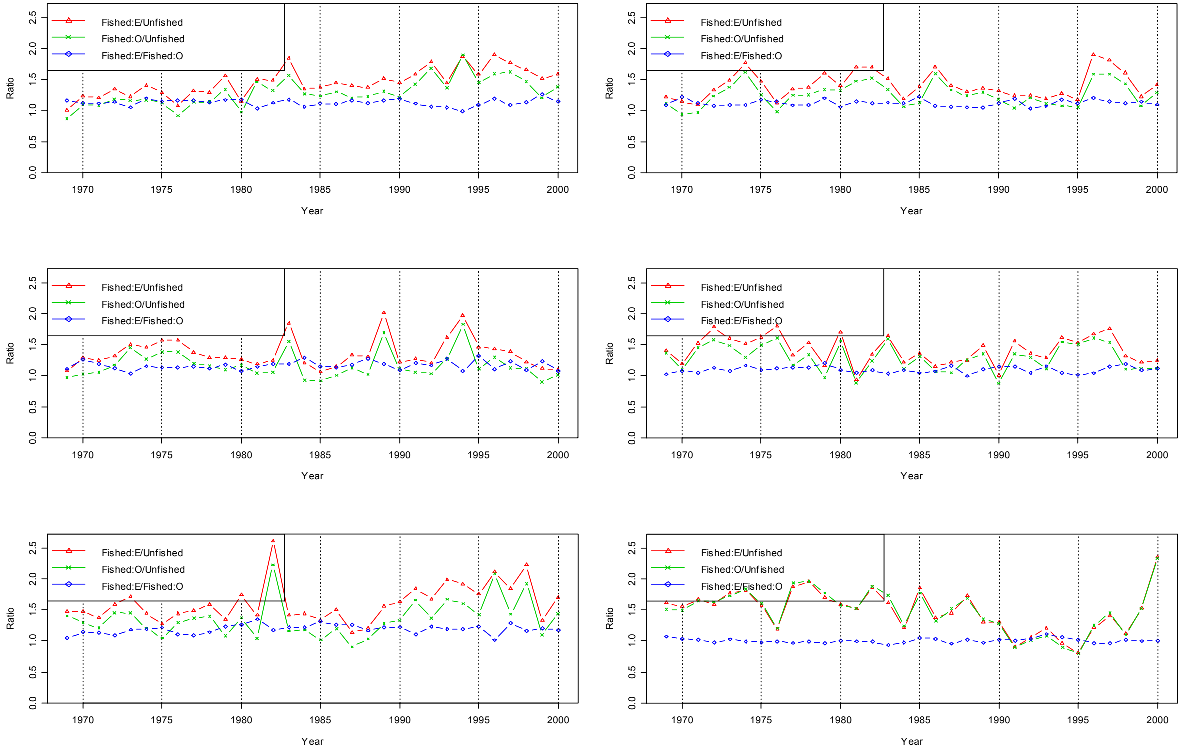

Figure 25: The ratio of CPUE for fished and unfished cells of five simulated test data sets when the spatial pattern of effort is unchanged, but CPUE is randomly permuted in space for those cells where there is effort. The last panel shows the same information for the original model and test data.

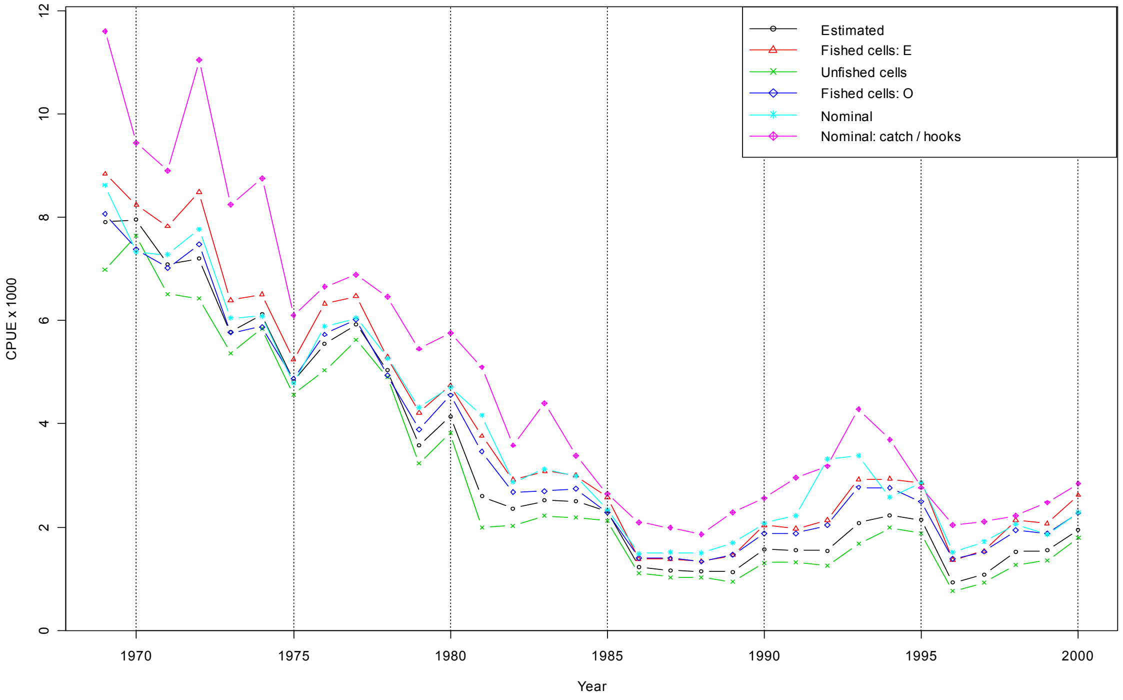

Figure 26: The non-weighted empirical and estimated bagging yearly CPUE for the fished and unfished cells when effort is randomly permuted in space and CPUE is then randomly permuted in space to those cells with effort.

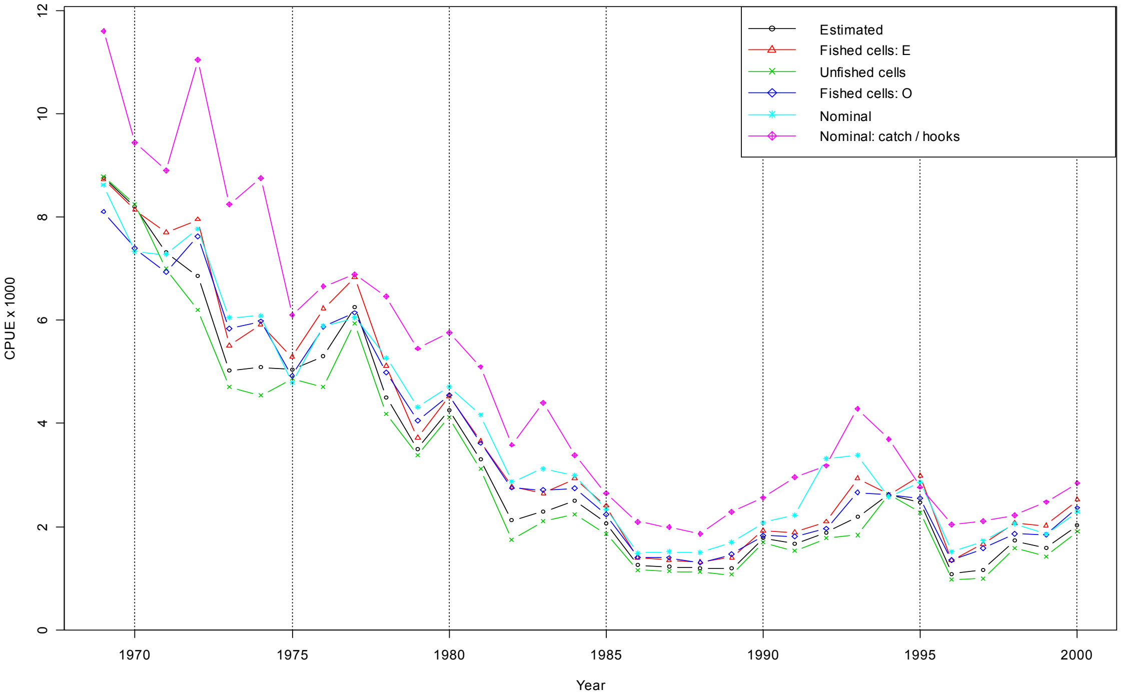

Figure 27: The non-weighted empirical and estimated bagging yearly CPUE for the fished and unfished cells when effort and CPUE are randomly permuted in space.

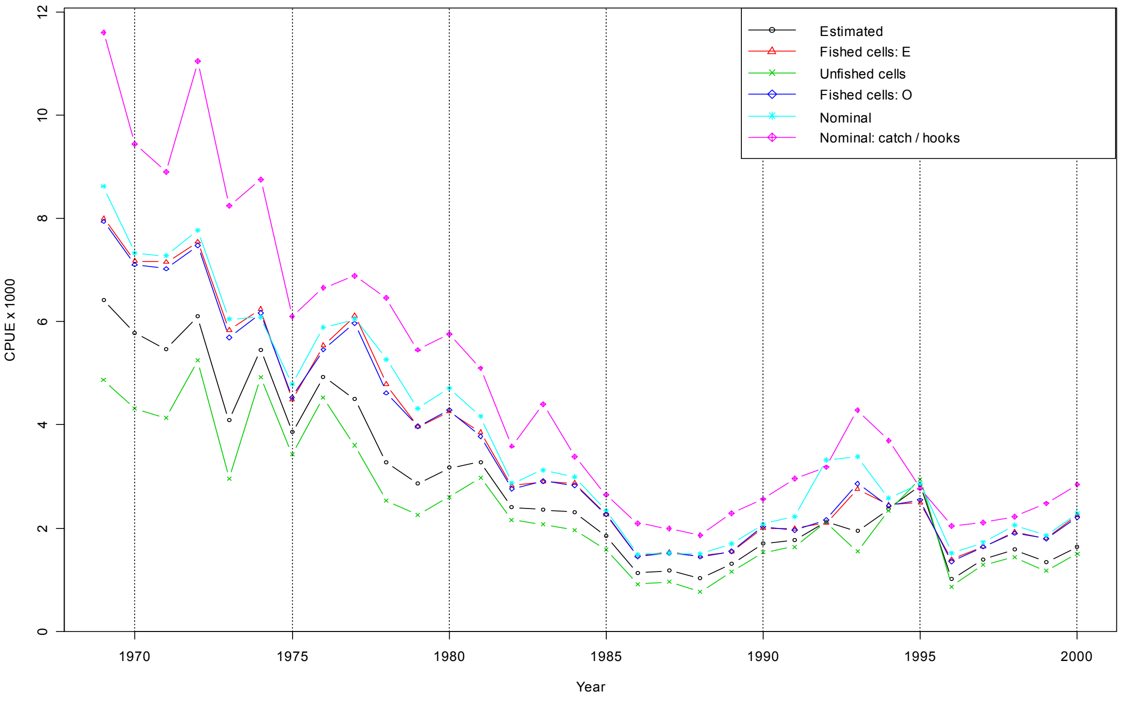

Figure 28: The non-weighted empirical and estimated bagging yearly CPUE for the fished and unfished cells when the spatial pattern for CPUE is unchanged, but effort is randomly permuted in space for those cells where there is effort.

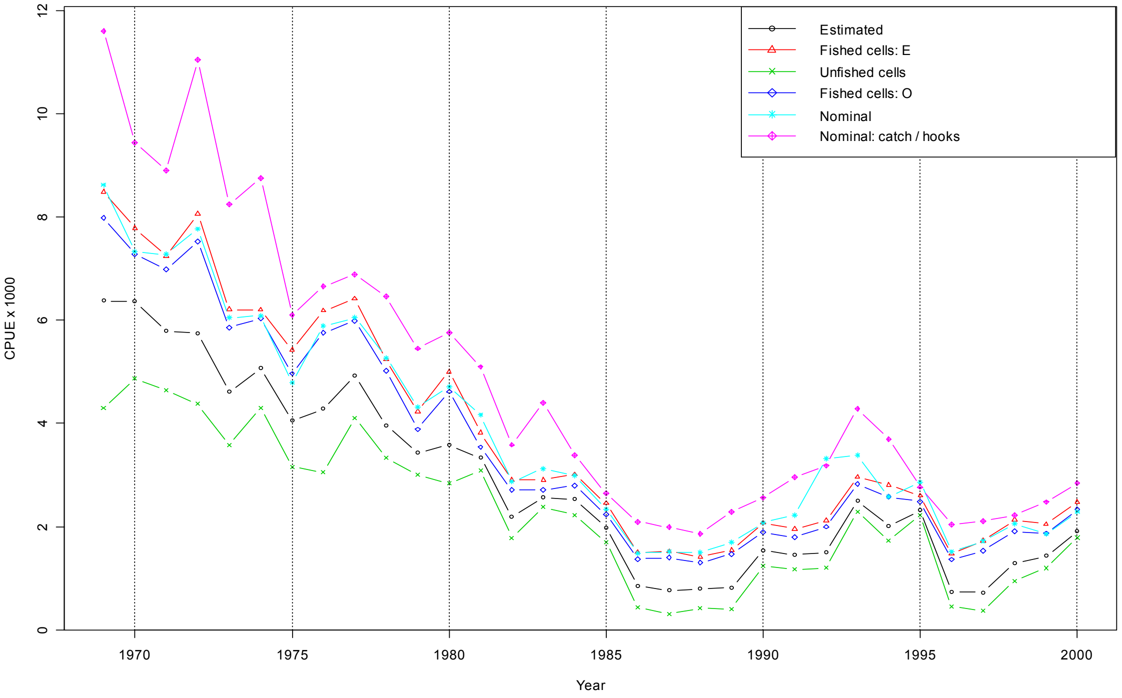

Figure 29: The non-weighted empirical and estimated bagging yearly CPUE for the fished and unfished cells when the spatial pattern of effort is unchanged, but CPUE is randomly permuted in space for those cells where there is positive effort.

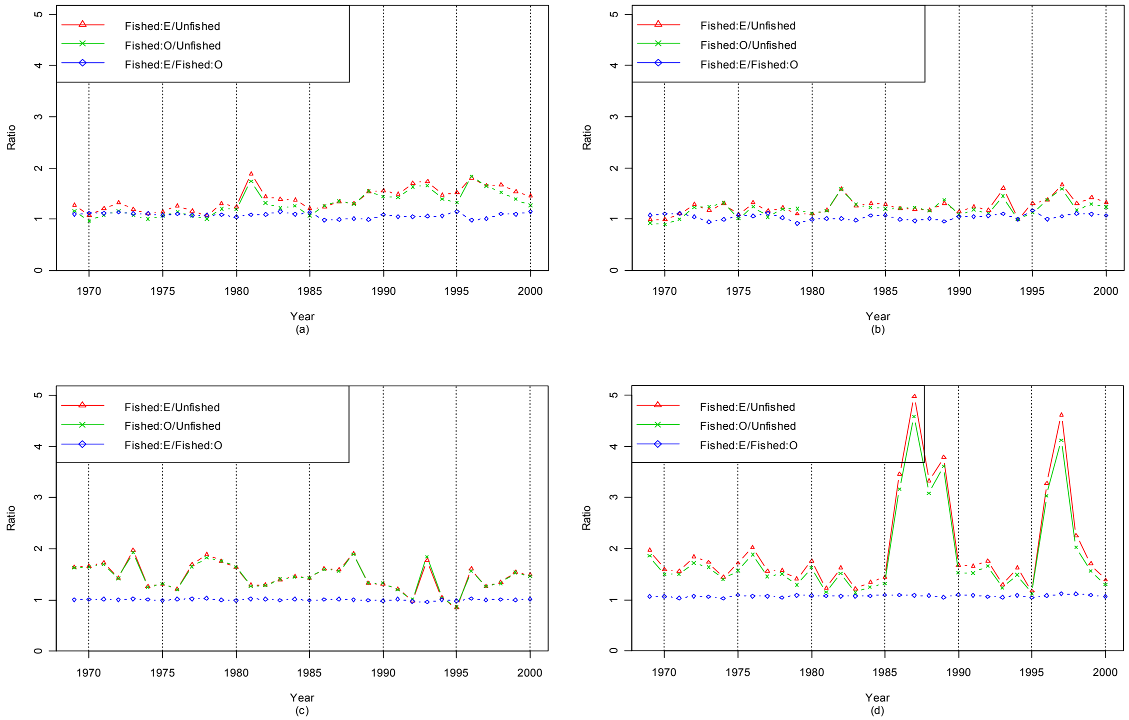

Figure 30: The ratio of non-weighted yearly CPUE for fished and unfished cells for (a) when effort is randomly permuted in space and CPUE is then randomly permuted in space to those cells with effort; (b) when effort and CPUE are randomly permuted in space; (c) when the spatial pattern of CPUE is unchanged, but effort is randomly permuted in space for those cells where there is effort; and (d) when the spatial pattern of effort is unchanged, but CPUE is randomly permuted in space for those cells where there is positive effort.